Learning Flexible Large Multimodal Models with Arbitrary Modality Combinations

Xinyu Zhao[1*], **Kangqi Ni**[1*], **Jie Peng**[1], **Ang Li**[2], **Tianlong Chen**[1†]
[1]University of North Carolina at Chapel Hill, [2]University of Maryland
{xinyu,kangqini}@cs.unc.edu

Abstract

Multimodal Large Language Models (MLLMs) show strong potential for cross-modal understanding by integrating powerful language models with multimodal encoders. However, extending MLLMs to handle a diverse range of modalities introduces two critical and intertwined challenges: (1) the reliance on fully paired multimodal data, often scarce or costly to acquire across all modalities, and (2) the computational inefficiency from processing numerous modality tokens and requiring substantial model updates for each new modality. To address these challenges, we enable MLLMs to handle missing modalities by generating representations for absent inputs. Furthermore, recognizing that an increasing number of modalities leads to linearly scaling token counts and that lengthy generated sequences can hinder performance, we employ a dual-stage compression mechanism. It first reduces the number of tokens per modality and then condenses information from multiple modalities into a single, compact token sequence. This culminates in Flex-M^3, a novel MLLM framework designed for flexible and efficient learning across arbitrary combinations of modalities. Experiments across diverse multimodal benchmarks and backbones demonstrate that Flex-M^3 robustly handles varied modality inputs and scales efficiently. Notably, Flex-M^3 outperforms its counterpart trained on only full-modality data, with consistent improvements of {2.29%, 3.15%, 11.01%} on multimodal reasoning tasks {NExT-QA, MUSIC-AVQA, SQA3D}. Moreover, Flex-M^3 model demonstrates superior robustness during inference, even when a high proportion of modalities are missing from the input samples. Our codes are released at https://github.com/UNITES-Lab/Flex-M3.

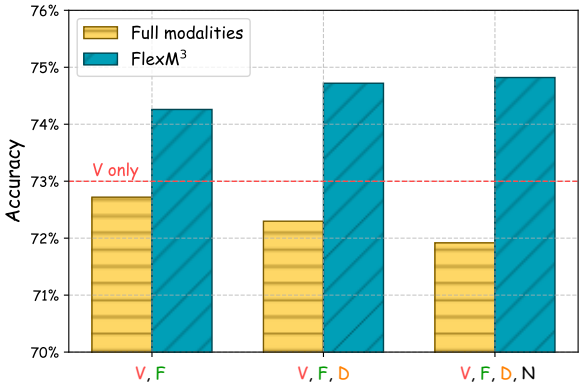

Figure 1: Comparison of accuracy (%) on a multimodal video question answering task NeXT-QA using different modality combinations for Flex-M^3 against a baseline trained on full modalities data only. The x-axis represents the available non-text modalities during fine-tuning: Video, optical Flow, Depth, and surface Normalization. The dashed red line indicates the performance when using only V.

1 Introduction

In recent years, Multimodal Large Language Models (MLLM) have become a popular paradigm in multimodal learning. MLLMs leverage the understanding and generative capabilities of pre-trained Large Language Models (Dubey et al., 2024; Achiam et al., 2023; Anil et al., 2023), enhancing them by integrating information from diverse perceptual inputs (*e.g.*, vision (Liu et al., 2024; Wang et al., 2024), speech (Zhang et al., 2023a; Chu et al., 2023), 3D (Xu et al., 2024), biomarker (Zhuo et al., 2024), and tabular information (Zheng et al., 2024)). Recent advancements are pushing towards omnipotent MLLMs managing numerous modalities to tackle complex scenarios, *i.e.* automated planning (Wei et al., 2024; Wang et al., 2023a) and world simulation (Ge et al., 2024).

However, realizing the full potential of MLLMs is challenged by data acquisition and training efficiency. Firstly, acquiring fully paired multimodal datasets is arduous. This could be attributed to real-world constraints, such as in biomedical settings where measurement devices might destroy paired

[*]Equal contribution
[†]Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 10664–10678

samples (Xi et al., 2024). Furthermore, collection costs vary drastically across modalities. For example, readily available image-text pairs are far more abundant than data for depth or thermal imaging (Zhu et al., 2024; Girdhar et al., 2023). Prior work has explored data synthesis, image translation (Bhat et al., 2023; Xu et al., 2023; Lee et al., 2023a), or meticulous training pipelines over disparate data resources (Han et al., 2024) to mitigate this. However, these methods often involve laborious data preparation and empirical tuning of training dynamics, limiting their generalizability.

A second critical challenge is the substantial computational cost associated with training and deploying MLLMs. Incorporating each new modality requires significant updates to the LLM to align textual representations with the new modal input. While research into efficient MLLMs proposes using separate projections or adapters to reduce trainable parameters (Li et al., 2023; Han et al., 2024; Yu et al., 2025), the inherent MLLM architecture that projects each modality into hundreds of tokens still leads to high training and inference costs. This is especially problematic with a growing number of modalities or computationally intensive modalities like video. Moreover, many efficient MLLMs lack flexibility, mandating the presence of all designated modalities, which restricts their use with a mixture of incomplete data.

In light of the above challenges, we posit that one critical next step for MLLMs reflecting real-world data scenarios is **"flexible multimodal learning"**, which is *enabling MLLMs to adeptly process diverse input samples, where each sample can present a different and potentially incomplete combination of available modalities*. To realize flexible multimodal learning, we introduce Flex-M^3 with a generation module synthesizing representations for any missing modalities by dynamically conditioning on the ones that are present. Then, we observed that the number of tokens, particularly those generated for missing inputs, significantly impacts training efficiency and final performance. As more modalities are introduced, this can lead to a linear scaling of tokens, and generating lengthy sequences for absent modalities can constrain performance. To mitigate this, Flex-M^3 incorporates a two-stage compression process. Initially, we compress the token representations from each modality encoder. Following that, all available modality representations, both those originally present and those newly generated, are further consolidated into a single, highly compact token sequence. This ensures that only the most salient and efficiently encoded cross-modal information is passed to the LLM. We validate the efficacy of Flex-M^3 across various MLLM backbones and diverse multimodal tasks. As illustrated in Figure 1, our approach not only robustly handles incomplete data but also achieves an average performance gain of nearly 3% compared to counterparts trained exclusively using full modality samples. This advantage becomes even more distinct in groups involving more modalities. In sum, the contributions of this study are four-fold:

- We formulate **flexible multimodal learning** for MLLMs, enabling learning on data samples with diverse and potentially incomplete modality combinations that reflect real-world data distributions.
- We propose Flex-M^3, a novel MLLM framework that dynamically synthesizes latent representations for absent modalities via a lightweight, prompt-conditioned generation module, enabling robust training and inference under arbitrary modality availability.
- We introduce a **two-stage compression** mechanism that first reduces per-modality token counts and then fuses all modality representations into a compact, fixed-length sequence, ensuring scalable and efficient LLM input regardless of modality count.
- Extensive experiments across diverse benchmarks (`NExT-QA`, `MUSIC-AVQA`, `SQA3D`) and backbones (BLIP-2, LLaVA) demonstrate that Flex-M^3 achieves consistent improvements (up to 11% on `SQA3D`) with minimal computational overhead.

2 Related Work

Recent advances in Multimodal Large Language Models (MLLMs) extend language models to process and reason over diverse perceptual inputs by aligning modality-specific encoders with language representations. Representative works such as BLIP-2 (Li et al., 2023) and LLaVA (Liu et al., 2024) introduce lightweight projection or query-based interfaces to bridge frozen encoders and LLMs, achieving strong performance on vision-language tasks with limited trainable parameters. Subsequent studies further expand MLLMs to additional modalities, including audio, depth, flow, and 3D information, through modular adapters or multimodal fusion mechanisms (Yu et al., 2025;

Zhang et al., 2023a). Despite these advances, existing MLLMs typically assume the presence of all designated modalities during both training and inference, limiting their applicability in realistic scenarios where modality availability is heterogeneous or incomplete.

Multimodal Learning with Missing Modalities. Robust learning under missing modalities has been widely studied in multimodal learning, motivated by practical challenges such as sensor failure, acquisition cost, and privacy constraints (Ma et al., 2022; Wei et al., 2023; Lee et al., 2023b; Qiu et al., 2023; Zhang et al., 2023c; Wu et al., 2024). Early approaches rely on heuristic imputation strategies (Parthasarathy and Sundaram, 2020; Zhang et al., 2020), while more recent learning-based methods recover missing information via data-level reconstruction or representation-level generation (Pham et al., 2019; Hoffman et al., 2016; Wang et al., 2023b). Representation-level approaches are generally more effective, as they capture cross-modal dependencies directly in the latent space (Zhou et al., 2021; Zhi et al., 2024). Recent works also explore architectural flexibility, such as mixture-of-experts routing or continual modality expansion (Yun et al., 2024; Yu et al., 2024). However, most existing methods focus on classification or recognition tasks and do not directly address large-scale multimodal language models, where missing-modality recovery must interact with token-based reasoning and generation.

Efficient Multimodal Large Language Models. To reduce the computational cost of MLLMs, a growing body of work investigates token-efficient multimodal representations. These methods include token pruning and merging (Chen et al., 2024; Shang et al., 2024), query-based resampling (Li et al., 2024b; Hu et al., 2024), complex compressor module (Zhang et al., 2025) and compact visual or video representations (Song et al., 2024; Maaz et al., 2024; Li et al., 2024a; Lin et al., 2023; Zhang et al., 2023b). Such approaches significantly improve training and inference efficiency, especially for video-based MLLMs. However, existing efficiency-oriented methods generally assume fully observed modalities and do not consider how generated or imputed modality representations should be integrated efficiently, and are largely orthogonal to the problem of flexible modality availability.

3 Methodology

Scaling MLLMs to arbitrary modality combinations requires jointly addressing two tightly coupled challenges: *recovering absent representations* and *preventing token explosion*. Our framework is grounded in three design principles. **(1)** A lightweight generation module synthesizes missing modality features from consistently available anchor modalities (*i.e.*, text and video), using shared generative prompts for modality-agnostic inductive bias and modality-specific MLPs to capture distinct target patterns. This design adds minimal overhead while enabling seamless plug-and-play compatibility with existing MLLMs. **(2)** A two-stage compression pipeline first removes intra-modality redundancy, particularly crucial for filtering noise in generated tokens, and then fuses all modality features into a fixed-length sequence, guaranteeing stable computational cost regardless of input modality count. **(3)** A reconstruction objective (Eq. 2) bridges the generation and compression stages, ensuring that synthesized representations faithfully approximate real modality features in the latent space. We detail each component below.

3.1 Preliminary

Multimodal Large Language Models (MLLMs) extend LLMs to process and reason over multiple modalities such as vision, speech, and 3D data. A typical MLLM consists of modality-specific encoders, an interfacing module, and an LLM. Each encoder $\mathcal{E}_m$ maps raw inputs $\mathbf{X}_m$ to high-level features $\mathbf{F}_m = \mathcal{E}_m(\mathbf{X}_m)$. The interfacing module $\mathcal{A}$ aligns these features with the LLM input space by projecting them into token sequences or embeddings, optionally using modality-specific learnable queries $\mathbf{Q}_m$ to distill salient information. The resulting modality tokens $\mathbf{H}_m = \mathcal{A}(\mathbf{F}_m, \mathbf{Q}_m)$ act as soft prompts (Li et al., 2023) that condition the LLM on multimodal context. The modality tokens $\mathbf{H}_m$ are then combined—typically concatenated or interleaved—with text embeddings $\mathbf{H}_t$ and fed into the LLM to generate outputs $\mathbf{Y}$. MLLMs are trained end-to-end using a language modeling objective, $\mathcal{L}_{\text{LM}} = -\sum_{t=1}^{T} \log P(y_t \mid y_{<t}, \mathbf{H}_m, \mathbf{H}_t; \theta)$, enabling joint multimodal reasoning.

3.2 Flexible Learning via Missing Modality Generation

To address the challenge of incomplete data modalities, where one or more modalities may be absent,

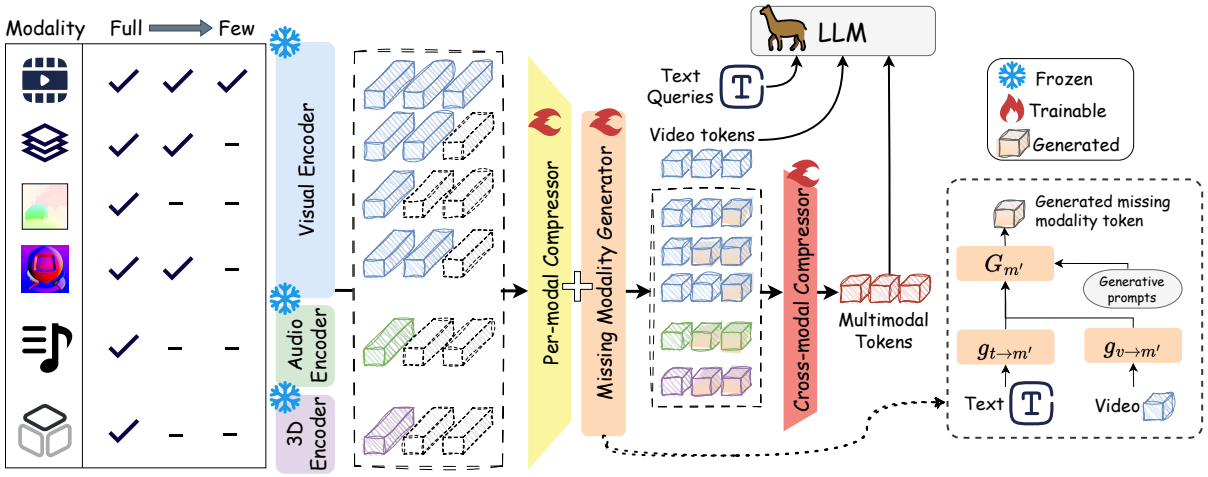

Figure 2: Overview of Flex-M^3 multimodal learning framework. The model processes arbitrary modality combinations by first generating missing modality embeddings using text and video-conditioned generative soft prompts (left subfigure). These, along with present modality tokens, undergo per-modal and then cross-modal compression to create a compact, robust representation. Finally, these compressed tokens, along with text and video, are input to an LLM.

we introduce a generation module to recover representations for missing modalities from the ones that are present. This approach allows the MLLM to effectively learn and operate across arbitrary combinations of available input data, significantly enhancing its flexibility.

The core of this generation process utilizes a set of consistently available modalities, *i.e.*, text and video inputs, as conditional information to recover other modalities termed "supportive" modalities. For each target "extra" modality m' (*e.g.*, depth, thermal, or other sensory data) that might be missing, we generate its feature representation. This generation is implemented by three components:

• **A learnable generative prompt $\mathbf{P}$**, a globally shared learnable tensor across all target missing modalities, which provides a modality-agnostic inductive bias for generation.

• **Modality-specific transformation networks**: For each target missing modality m', dedicated mapping functions $g_{t\to m'}(\cdot)$ and $g_{v\to m'}(\cdot)$, implemented as modality-specific MLPs, project the text embedding $\mathbf{H}_t$ and visual embedding $\mathbf{H}_v$ into a common hidden dimension d. Crucially, generating different modalities (*e.g.*, depth vs. flow) uses entirely different MLPs, ensuring each transformation captures the distinct characteristics of its target modality.

• **Final generation network**: The projected features from the transformation networks are concatenated with the generative prompt $\mathbf{P}$ along the sequence length dimension, and then processed by a dedicated generation MLP $G_{m'}(\cdot)$, which maps the concatenated sequence into N_q modality tokens to produce the final synthesized embedding $\hat{\mathbf{H}}_{m'}$.

In sum, the generation process for a missing modality m' can be formulated as Equation 1, where concat denotes the concatenation operation along the sequence dimension.

$$\hat{\mathbf{H}}_{m'} = G_{m'}\left(\text{concat}\left(\mathbf{P}, g_{t\to m'}(\mathbf{H}_t), g_{v\to m'}(\mathbf{H}_v)\right)\right) \quad (1)$$

This architecture allows for the generation of multiple missing modalities, using the same set of source modalities and the shared generative prompt, but with distinct and lightweight transformation procedures. The generation modules are trained end-to-end with the rest of the MLLM. To enable learning on generating high quality missing modality embeddings, we employ a reconstruction objective. During training, for data samples where a modality m is physically present, we stochastically treat it as "missing". In such cases, we obtain a generated feature $\hat{\mathbf{H}}_m$. Then, we compute a reconstruction loss, typically the Mean Squared Error (MSE), between the generated features $\hat{\mathbf{H}}_m$ and the presented real features $\mathbf{H}_m$. This loss is formulated as in Equation 2. The overall training objective for the MLLM is a combination of the standard language modeling loss and the weighted reconstruction losses: $\mathcal{L} = \mathcal{L}_{\text{LM}} + \lambda\mathcal{L}_{\text{Rec}}$, where λ is the weighting factor for reconstruction loss.

$$\mathcal{L}_{\text{Rec}} = \frac{1}{n}\sum_{m=1}^{n}\frac{1}{D}\sum_{i=1}^{D} \parallel \hat{h}_m^i - h_m^i \parallel \quad (2)$$

3.3 Modality Token Compression for Robust Generation

While the integration of multiple modalities enriches the context, the direct concatenation of all modality tokens can lead to a prohibitively large number of input tokens for the LLM. This not only escalates computational cost but can also introduce noise or redundant information, potentially hampering the robustness of the synthesized outputs. To mitigate these issues, we employ a two-stage strategy for compressing and refining modality tokens before they are processed by the main LLM. This strategy involves per-modality token compression and cross-modal token compression.

Per-Modal Compression The initial projected feature representations for each modality m, denoted as $\mathbf{H}_m$, and including any generated features $\hat{\mathbf{H}}_{m'}$ are often lengthy. To reduce this length, we apply a per-modality compression module, $\mathcal{C}_m$. It is designed to distill the most salient information from $\mathbf{H}_m$ into a more compact representation, $\mathbf{H}_m^{(c)}$. The compression module employs a set of N_q learnable query embeddings, *e.g.* $q_m \in \mathbb{R}^{N_q \times d}$ for modality m, where d is the dimension of query embedding and N_q is significantly smaller than the original token length of $\mathbf{H}_m$. These queries interact with the input modality tokens through the cross-attention mechanism. For the missing modalities, we switch to generate the per-modal compressor output $\hat{\mathbf{H}}_{m'}^{(c)}$.

Cross-Modal Compression After per-modality condensation, concatenating the resulting tokens with text embeddings still leads to a long input sequence for the LLM, especially when there are more modalities. To further condense the input and enable earlier cross-modal interactions, we introduce a cross-modal compression stage. This stage creates a more integrated and compact set of supportive modality tokens before interacting with the LLM. In this stage, all compressed modality tokens are concatenated and then processed by a cross-modal compression module $f(\cdot)$, generating a fused multimodal representation with fixed length for any number of input modalities M as:

$$\mathbf{Z} = f(\texttt{concat}(\mathbf{H}_0^{(c)}, \mathbf{H}_1^{(c)}, \ldots, \mathbf{H}_M^{(c)})) \quad (3)$$

The cross-modal compression output $\mathbf{Z}$, along with visual and text embeddings, are finally presented to the main LLM. This two-stage compression approach not only reduces the computational burden on the LLM but also aims to improve the robustness of generation by enabling the model to focus on the most salient cross-modal information, effectively filtering redundancies and noise.

4 Experiment

4.1 Experiment Setup

Datasets Details We evaluate Flex-M^3 on the 3 multimodal video reasoning and QA tasks: `NExT-QA`, `SQA3D`, and `MUSIC-AVQA`, which are detailed in Appendix A.2. We incorporate optical flow, depth maps, and surface normals extracted from the videos as additional modalities to enhance the model's understanding as Yu et al. (2025). Specifically, ZoeDepth (Bhat et al., 2023), Unimatch (Xu et al., 2023), and NLL-AngMF (Bae et al., 2021) are employed to extract depth, flow, and normal modalities.

Model Implementation and Training setup. For **pretrained modality encoders**, we utilize ViT-G (Sun et al., 2023) for all visual modalities including videos, depth, norm and flow. For non-visual modalities, we use BEATs (Chen et al., 2023) as the encoder for audio, and extract 3D point cloud features offline following 3D-LLM (Hong et al., 2023a). For **MLLM model backbone**, we implement Flex-M^3 on BLIP-2 (Li et al., 2023) and LLaVA (Liu et al., 2024) to showcase Flex-M^3's general applicability. Detailed training hyperparameters are shown in Appendix A.2. The entire model is trained end-to-end with the standard language modeling loss and an auxiliary generation reconstruction loss with weight $\lambda = 0.001$.

• We adapt BLIP-2's initial Q-Former architecture as per-modal compressor, with query token number $N_q = 32$. The cross-modal compressor is implemented as a modality-specific linear layer that projects the output features from the corresponding Q-Former into the language model. For fine-tuning, we initialize Flex-M^3 from BLIP-2 where the encoders and LLM are frozen, and only the per-modal compressors, cross-modal compressor, and generator are updated. To further enhance fine-tuning efficiency, we update per-modal compressors using LoRA (Hu et al., 2022) with rank 64.

• For LLaVA-based Flex-M^3, we similarly integrate our generation and compression mechanisms with Llama 3.1 8B language model and ViT-L visual processing pipeline, following (Liu et al., 2024). Similar to the settings in BLIP-2, we initialize LLaVA with pretrained per-modal and cross-

modal compressor from LLaVA-Mini (Zhang et al., 2025), where we copy the compressors for modalities other than video. The per-modal compressor is a 2D perceiver-resampler network with 8×8 learnable queries as input, while the cross-modal compression module is a 4-layer Transformer decoder. We finetune Flex-M^3 for all model components on LLaVA except for the encoders, as we find the performance gain after enabling the language model to be updated is significant while the computation cost growth is moderate.

Baselines and Evaluation Setup To support the effectiveness of Flex-M^3, we consider three groups of comparison baselines: (1) **Essential Modalities Only**: These models utilize the full dataset but are restricted to processing only the essential text and video modalities. This baseline is also evaluated on text and video modalities only. (2) **Full Modalities with Incomplete Data**: This baseline illustrates the data inefficiency of standard MLLMs that require a fixed, complete set of input modalities. Without a missing-modality mechanism, such models can only be trained on samples where all M modalities are simultaneously present. To simulate realistic heterogeneous modality availability, the original dataset is divided into 2^M subsets, each corresponding to one possible modality combination and containing $1/2^M$ of the original data volume. For example, with one additional modality (e.g. Video, Flow in the first experiment group of NExT-QA) in Table 1, 50% of the non-text samples contain V only while the rest contain all modalities. This baseline is evaluated on full modalities. (3) **Learnable Padding for Missing Modalities**: This baseline employs the full dataset while accommodating arbitrary modality combinations through a learnable padding technique. Specifically, [PAD] tokens from the LLM embedding space, are used to represent absent modalities. These padded inputs are then processed by the cross-modal compressor, enabling fusion of the padding with existing modalities. This improved baseline and Flex-M^3 are evaluated on full modalities.

4.2 Main Results

Superior Performance of Flex-M^3 with Flexible Modality Learning The fine-tuning results on NExT-QA, presented in Table 1, compellingly demonstrate that Flex-M^3 excels in handling various multimodal inputs, particularly in scenarios characterized by missing modalities. Taking Flex-M^3 with BLIP-2 as example, ❶ when utilizing the full dataset with arbitrary modality combinations (indicated by "Missing: ✓"), Flex-M^3 consistently outperforms alternative approaches. For instance, in the V, F, D, N setting, Flex-M^3 achieves an average score of 74.82, surpassing both the "Padding" baseline (74.22) and the "Essential Modalities Only" baseline (V: Avg. 73.00). This highlights Flex-M^3's proficiency in leveraging supportive information from additional modalities, even when their presence is not guaranteed. ❷ This contrasts sharply with the "Full Modalities with Incomplete Data" baseline (rows marked "Full w/ Inc. Data"), which exhibits a performance decline as more modalities are introduced (from 73.00 for V only, down to 72.04 for V, F, D, N). This performance drop could be attributable to the MLLM being fine-tuned on progressively smaller, specific data subsets for each modality combination ($1/2^M$ of the original data volume), which hampers generalization. ❸ Flex-M^3 not only overcomes this limitation but also consistently betters the "Padding" method across all tested auxiliary modality counts: achieving a $+1.00$ point gain with one auxiliary modality (V, F: Flex-M^3 74.26 vs. Padding 73.26) and a $+0.60$ point gain with three (V, F, D, N: Flex-M^3 74.82 vs. Padding 74.22). This sustained advantage is attributed to Flex-M^3's modality-specific generation and compression design, which effectively distills key information and manages modality absence more adeptly than simple learnable padding. ❹ Furthermore, this robust performance extends across diverse question categories (Causal, Temporal, Descriptive Average Performance), where Flex-M^3 generally secures the highest scores in settings with multiple potential modalities. In essence, Flex-M^3 showcases a significant capability in flexibly and efficiently integrating information from an arbitrary set of available modalities, underscoring the efficacy of its advanced modality compression techniques for robust multimodal understanding in the face of incomplete data.

Generalization of Flex-M^3 across Different MLLM Backbones To further substantiate the generalizability of Flex-M^3, we evaluated its efficacy when integrated with LLaVA architecture (Liu et al., 2024). The results presented in Table 1 (bottom), again validate the effectiveness of Flex-M^3 against strong video-LLMs fine-tuned with extra supportive modalities. Flex-M^3 with LLaVA demonstrate a substantial average performance increase of approximately 10.83% points compared

Table 1: Performance on Video Question Answering (NExT-QA). Notations for each modality and question type are: **V**: Video RGB frames, **F**: optical Flow, **D**: Depth, and **N**: surface Normalization. **P.&N.**: Prev & Next, **Pre.**: Present, **Cnt.**: Count, **Loc.**: Location, and **Otr.**: Other. Method names: **Essential** uses only text and video; **Full w/ Inc. Data** trains a standard MLLM on all modalities with incomplete data splits ($1/2^M$ per combination); **Padding** replaces missing modalities with learnable tokens. Within each group, the best result is **bold** and the second-best is underlined. All results are percentages.

Modality	Missing	Method	Causal			Temporal			Descriptive				Avg.
			How	Why	Avg.	P.&N.	Pre.	Avg.	Cnt.	Loc.	Otr.	Avg.	
BLIP-2													
V	✗	Essential	69.69	74.64	73.34	65.14	72.55	74.84	64.41	92.20	81.31	81.60	73.00
V, F	✗	Full w/ Inc. Data	69.55	74.43	73.15	64.58	72.85	74.20	66.10	91.53	79.67	81.08	72.72
	✓	Padding	**71.16**	74.58	73.69	65.25	73.00	74.97	64.97	**92.54**	81.64	81.98	73.26
	✓	Flex-M^3	70.42	**75.99**	**74.53**	**66.93**	**73.60**	**76.89**	**65.54**	92.20	**83.28**	**82.63**	**74.26**
V, F, D	✗	Full w/ Inc. Data	66.91	73.86	72.04	64.25	72.85	73.81	**64.97**	93.56	80.66	81.98	72.30
	✓	Padding	70.28	**76.20**	74.65	66.03	**73.45**	75.87	64.41	**93.90**	81.64	**82.37**	74.02
	✓	Flex-M^3	**73.06**	75.68	**74.99**	**67.15**	74.36	**77.15**	63.84	91.53	**83.61**	82.11	**74.72**
V, F, D, N	✗	Full w/ Inc. Data	66.91	73.86	72.04	64.25	72.85	73.81	**64.97**	93.56	80.66	81.98	72.30
	✓	Padding	**72.62**	75.31	74.61	66.59	74.51	76.51	63.28	93.22	81.97	81.98	74.22
	✓	Flex-M^3	70.66	**76.62**	**75.06**	**66.85**	74.81	**76.84**	64.41	**93.90**	**84.92**	**83.66**	**74.82**
LLaVA													
V	✗	Essential	74.38	77.23	76.49	68.60	75.17	71.53	59.32	93.22	84.92	82.24	75.78
V, F	✗	Full w/ Inc. Data	71.89	76.30	75.14	68.16	74.34	70.91	58.76	92.88	85.57	82.24	74.88
	✓	Padding	76.28	77.39	77.10	**69.39**	75.45	**72.08**	61.58	92.54	**86.23**	83.01	76.40
	✓	Flex-M^3	**77.01**	**77.96**	**77.71**	68.04	**75.87**	71.53	**62.15**	**93.22**	**86.23**	**83.40**	**76.60**
V, F, D	✗	Full w/ Inc. Data	64.71	69.13	67.97	61.45	66.53	63.71	55.37	88.81	78.36	77.09	68.01
	✓	Padding	75.7	77.34	76.91	70.39	**77.55**	**73.57**	**61.58**	92.54	**86.23**	**83.01**	76.78
	✓	Flex-M^3	**77.89**	**78.33**	**78.21**	**71.28**	75.17	73.01	60.45	**92.88**	**86.23**	82.88	**77.26**
V, F, D, N	✗	Full w/ Inc. Data	49.63	54.05	52.90	49.83	55.23	52.23	51.98	80.68	64.26	67.70	54.98
	✓	Padding	74.38	**78.22**	77.22	68.27	76.15	71.77	62.71	92.2	**88.85**	**84.17**	76.56
	✓	Flex-M^3	**77.89**	77.91	**77.91**	**70.84**	**76.43**	**73.33**	**63.84**	**92.54**	86.56	83.66	**77.32**

to training with full modality samples only. This consistent improvement demonstrates that the architectural benefits of Flex-M^3 can be effectively transferred across foundational models. Importantly, the generation and compression modules of Flex-M^3 are plug-and-play components that operate between the modality encoders and the LLM, and can be integrated into any MLLM following the standard "encoder → connector → LLM" pipeline without modifying its pretrained weights.

Generalization of Flex-M^3 across Non-visual modalities To further evaluate whether Flex-M^3 can extend to non-visual modalities that the model backbone has not been pre-trained or fine-tuned on, we perform fine-tuning and evaluation on the MUSIC-AVQA and SQA3D benchmarks. Results are presented in Table 2 and Table 3. ❶ On the MUSIC-AVQA benchmark, Flex-M^3 demonstrates its surprising capacity for audio-video reasoning. When leveraging auxiliary modality information where samples contain missing modalities, Flex-M^3 achieves over 11% improvement over the baseline learned on full-modality data only. Also, compared to the baseline finetuned on text-video modalities, Flex-M^3 obtains performance gains comprehensively across all question subclasses (audio, visual, audio-visual). This again validates the benefit of utilizing diverse modality combinations, and the potential of flexible multimodal learning. ❷ The results in SQA3D again validate the versatility and effectiveness of Flex-M^3, where it achieves the leading average accuracy of 53.48% (+3.15% over the full-modality data baseline). 3D-associated video reasoning tasks require a model to interpret dynamic visual narratives from video with static and rich spatial, geometric information from 3D modalities. The ability of Flex-M^3 to leverage these combined inputs allows it to construct a more holistic and nuanced understanding of the scene. Computation analysis of Flex-M^3 is provided in Appendix A.1.

Robustness to Missing Modalities at Inference While from the evaluation with full modalities in Table 1-3, both padding and Flex-M^3 outperform other baselines, the distinction emerges when assessing their performance under random modality absence during inference. We take NExT-QA with V, F, D, N modalities as example. We randomize missing conditions for each sample, where 1

Table 2: Performance on Audio-Video Question Answering (MUSIC-AVQA) with BLIP-2-based Flex-M^3 and baseline models. Notations for each modality and question type are: **V**: *Video RGB frames*, **A**: *Audio*, **F**: *optical Flow*, **D**: *Depth*, and **N**: *surface Normalization*. **Cnt.**: *Counting*, **Com.**: *Comparative*, **Loc.**: *Location*, **Ext.**: *Existential*, and **Tem.**: *Temporal*. Method names: **Essential** uses only text and video; **Full w/ Inc. Data** trains a standard MLLM on all modalities with incomplete data splits; **Padding** replaces missing modalities with learnable tokens. Within each group, the best result is **bold** and the second-best is underlined. All results are percentages (%).

Modality	Missing	Method	Audio			Visual			Audio-Visual						Avg.
			Cnt.	Com.	Avg.	Cnt.	Loc.	Avg.	Cnt.	Ext.	Loc.	Com.	Tem.	Avg.	
V	✗	Essential	88.14	60.73	82.21	85.73	87.11	86.40	82.93	84.34	69.66	62.35	73.04	74.65	76.28
V, A, F, D, N	✗	Full w/ Inc. Data	79.75	57.09	74.85	75.75	77.05	76.38	69.65	80.54	58.71	56.38	68.18	66.79	70.93
	✓	Padding	89.49	**65.18**	**84.22**	**87.03**	90.43	88.69	**85.67**	83.11	71.49	**67.59**	73.04	76.54	81.17
	✓	Flex-M^3	**89.71**	62.75	83.87	**87.03**	**92.48**	**89.69**	84.93	**85.12**	**73.74**	66.87	**74.14**	**77.19**	**81.94**

Table 3: Performance on Situated Question Answering (SQA3D) with BLIP-2-based Flex-M^3 and baseline models. Notations for each modality and question type are: *Video RGB frames*, **V**: *Bird-Eye View image*, **P**: *3D Point cloud*, **D**: *Depth*, and **N**: *surface Normalization*. Method names: **Essential** uses only text and video; **Full w/ Inc. Data** trains a standard MLLM on all modalities with incomplete data splits; **Padding** replaces missing modalities with learnable tokens. Within each group, the best result is **bold** and the second-best is underlined. All results are percentages (%).

Modality	Missing	Method	What	Is	How	Can	Which	Others	Avg.
V	✗	Essential	44.99	47.74	63.02	64.88	47.59	49.11	51.69
V, P, D, N	✗	Full w/ Inc. Data	43.59	45.38	63.02	59.97	50.42	49.29	50.33
	✓	Padding	45.86	45.81	65.98	**65.95**	**49.86**	**54.26**	53.25
	✓	Flex-M^3	**46.82**	**47.74**	**66.57**	**65.95**	47.03	53.55	**53.48**

to 3 supportive modalities (from F, D, N) could be absent. We use the Missing Ratio (MR) to denote the overall proportion of missing modalities across the entire test set. As depicted in Figure 3d, the performance of naive padding approaches degrades as the MR increases, a trend observed across both LLaVA and BLIP-2 based models. In contrast, Flex-M^3, leveraging modality-specific generation, exhibits robust performance in both settings. The accuracy of Flex-M^3 models remains stable or even slightly increases under high MR situations (70%), consistently outperforming the padding counterparts. This underscores a key advantage of Flex-M^3. While naive padding falters with substantial data incompleteness at inference, Flex-M^3 can manage modality variations through generation, providing a more resilient framework for MML.

4.3 Extra Analysis and Ablation Studies

To identify the optimal design of Flex-M^3, we analyze its module contributions, hyperparameter sensitivity, and training efficiency. All experiments are conducted with the BLIP-2 backbone on the 10% NExT-QA subset, trained for 5 epochs using all supportive modalities. More ablations of Flex-M^3 components are detailed in Appendix A.3.

Ablation on Generation Loss Weight. An appropriate choice of the generation loss weight could benefit the performance of Flex-M^3. We compare Flex-M^3 under different generation loss weights (λ) in Figure 3a. The results indicate that a moderate weight ($1e^{-3}$) appears to yield optimal accuracy. Performance drops noticeably when λ is either significantly lower or higher. This suggests that while the reconstruction loss is crucial for learning to recover missing modalities, its contribution must be carefully balanced against the primary language modeling objective to prevent it from interfering with the core task.

Generation token numbers. In Figure 3b, we study the impact of the number of generation tokens (N_q) for all modalities on both accuracy and computational cost. As the token number increases to 32, accuracy generally improves. However, further increasing N_q to 64 results in a slight decrease in accuracy. This suggests that $N_q = 32$ reaches an optimal balance between representational capacity for the generated tokens and computational efficiency, with larger values potentially introducing redundancy. Moreover, we investigate altering the generation token numbers across modalities while keeping the token numbers for other modalities fixed at 32. The results in Figure 3c highlight

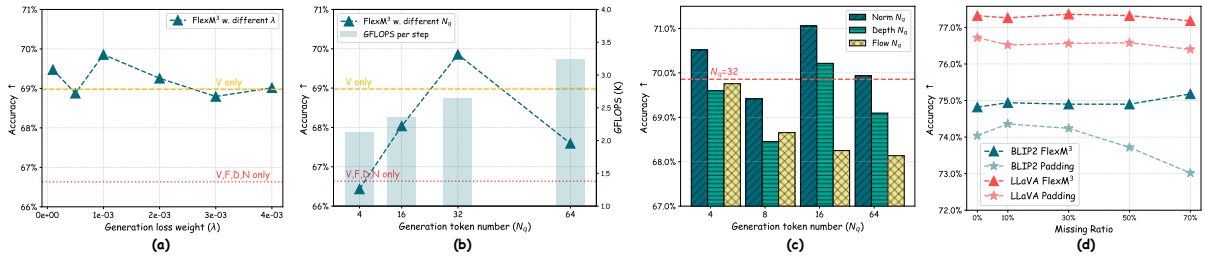

Figure 3: Extra studies on Flex-M^3 hyperparameters. **(a)** investigates the effect of varying the generation loss weight (λ) on model performance. **(b)** examines the impact of different generation token numbers (N_g) on accuracy and computational cost (GFLOPS per training step). **(c)** compares the impact of separately changing generation tokens per modality while keeping other modalities N_q=32. All experiments are conducted on NeXT-QA with V, F, D, N modalities. **(d) Comparison between BLIP-2 and LLaVA-based Flex-M^3 and Padding baseline on random modality missing evaluation.**

Table 4: Cosine similarity between generated and real modality embeddings on three benchmarks (BLIP-2). Higher values indicate better generation quality. Flex-M^3 produces embeddings highly aligned with real features, while Padding and Random yield near-zero similarity.

Dataset	Modality	Flex-M^3	Padding	Random
NExT-QA	Norm	0.8911	0.0138	0.0017
	Depth	0.8817	0.0194	−0.0019
	Flow	0.8914	0.0320	−0.0041
	Avg.	**0.8881**	0.0217	−0.0014
MUSIC-AVQA	Norm	0.5795	0.0090	−0.0050
	Depth	0.7465	0.0305	−0.0002
	Flow	0.5773	−0.0299	−0.0047
	Audio	0.5114	0.0065	−0.0023
	Avg.	**0.6344**	0.0032	−0.0033
SQA3D	Norm	0.9631	−0.0094	−0.0037
	Depth	0.5139	−0.0213	0.0012
	PC	0.9120	0.0218	−0.0003
	Avg.	**0.7963**	−0.0030	−0.0009

how individual modalities could benefit from different representational capacities during generation. Overall, $N_q = 32$ achieves moderately high accuracy for all modalities, and more tokens do not guarantee higher performance, aligning with previous findings in Figure 3b. Interestingly, some modalities could even improve with smaller N_q. For example, $N_q = 16$ yields better results for the Norm and Depth modalities. These findings suggest that we could dynamically adjust the generation token number per modality for flexible multimodal learning.

Quality of generated representations. A natural question is whether the generation module produces semantically meaningful embeddings or merely acts as a training regularizer. To answer this, we sample 100 test examples per benchmark and compute the cosine similarity between generated and real modality embeddings. As reported in Table 4, Flex-M^3 achieves high alignment with real features across all benchmarks (average similarity of 0.89, 0.63, and 0.80 on NExT-QA, MUSIC-AVQA, and SQA3D, respectively), while both Padding and Random baselines yield near-zero similarity. This confirms that the performance gains of Flex-M^3 stem from faithful modality-consistent reconstruction in the latent space, rather than trivial token filling. Three additional observations reinforce this conclusion. First, the Padding baseline is also trained on diverse modality combinations with learnable tokens, so it shares the same regularization benefit of exposure to incomplete inputs; the consistent performance gap over Padding (Tables 1–3) is therefore attributable to semantic generation. Second, as shown in Figure 3d, if the generator were purely a regularizer, its benefit would be fixed after training and would diminish under increasing modality absence at inference. Instead, Flex-M^3 maintains or slightly improves accuracy even at 50–70% missing ratios, while Padding degrades sharply. Third, the cross-dataset variation in similarity (e.g., higher on NExT-QA than MUSIC-AVQA) correlates with task-level modality complementarity, suggesting the generator adapts to how much each auxiliary modality contributes to the reasoning task. We further compare Flex-M^3 against the missing modality bank from Flex-MoE (Yun et al., 2024) in Appendix A.4, where Flex-M^3 outperforms it by a substantial margin, confirming the advantage of instance-conditioned generation over static global token banks.

5 Conclusion

Existing multimodal MLLMs necessitate complete sets of modality inputs for training and inference, limiting their ability to utilize the prevalent heterogeneous and incomplete multimodal data. This paper introduced Flex-M^3, a novel MLLM designed to adeptly process data featuring arbitrary combinations of modalities. Extensive experiments demonstrate that Flex-M^3 achieves significant performance gains across various MLLM backbones and diverse multimodal benchmarks, with minimal additional computational overhead.

Limitations

While Flex-M^3 shows strong flexibility and robustness incorporating arbitrary modality combinations, some limitations remain. First, it takes at least one or two consistently available anchor modalities (e.g., text and video) to condition the generation of missing modalities. Scenarios where all informative modalities are simultaneously absent, or where the remaining modalities are extremely weak, are beyond the current scope. Second, although the proposed generation and compression modules are a lightweight solution compared to full model retraining, they still introduce additional components and hyperparameters that require tuning for optimal performance. Finally, our evaluation focuses on multimodal reasoning benchmarks with structured auxiliary modalities; extending Flex-M^3 to open-ended or highly noisy real-world modalities, as well as to continual or streaming modality arrival settings, is an important future direction.

Acknowledgments

This work was supported in part by the National Institutes of Health (NIH) under award 1R01EB037101-01. The opinions and conclusions expressed herein are those of the authors and do not necessarily reflect the official policies or positions, either expressed or implied, of the NIH.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Gwangbin Bae, Ignas Budvytis, and Roberto Cipolla. 2021. Estimating and exploiting the aleatoric uncertainty in surface normal estimation. In *Proceedings of the International Conference on Computer Vision (ICCV)*, pages 13137–13146.

Shariq Farooq Bhat, Reiner Birkl, Diana Wofk, Peter Wonka, and Matthias Müller. 2023. Zoedepth: Zero-shot transfer by combining relative and metric depth. *CoRR*, abs/2302.12288.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 19–35. Springer.

Sanyuan Chen, Yu Wu, Chengyi Wang, Shujie Liu, Daniel Tompkins, Zhuo Chen, Wanxiang Che, Xiangzhan Yu, and Furu Wei. 2023. BEATs: Audio pre-training with acoustic tokenizers. In *Proceedings of the International Conference on Machine Learning (ICML)*.

Yunfei Chu, Jin Xu, Xiaohuan Zhou, Qian Yang, Shiliang Zhang, Zhijie Yan, Chang Zhou, and Jingren Zhou. 2023. Qwen-audio: Advancing universal audio understanding via unified large-scale audio-language models. *arXiv preprint arXiv:2311.07919*.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Zhiqi Ge, Hongzhe Huang, Mingze Zhou, Juncheng Li, Guoming Wang, Siliang Tang, and Yueting Zhuang. 2024. WorldGPT: Empowering LLM as multimodal world model. In *ACM Multimedia 2024*.

Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. Imagebind one embedding space to bind them all. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 15180–15190. IEEE.

Jiaming Han, Kaixiong Gong, Yiyuan Zhang, Jiaqi Wang, Kaipeng Zhang, Dahua Lin, Yu Qiao, Peng Gao, and Xiangyu Yue. 2024. Onellm: One framework to align all modalities with language. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26584–26595.

Judy Hoffman, Saurabh Gupta, and Trevor Darrell. 2016. Learning with side information through modality hallucination. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 826–834.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023a. 3d-llm: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Yining Hong, Haoyu Zhen, Peihao Chen, Shuhong Zheng, Yilun Du, Zhenfang Chen, and Chuang Gan. 2023b. 3d-LLM: Injecting the 3d world into large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation

of large language models. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Wenbo Hu, Zi-Yi Dou, Liunian Harold Li, Amita Kamath, Nanyun Peng, and Kai-Wei Chang. 2024. Matryoshka query transformer for large vision-language models. *Preprint*, arXiv:2405.19315.

Dong-Guw Lee, Myung-Hwan Jeon, Younggun Cho, and Ayoung Kim. 2023a. Edge-guided multi-domain rgb-to-tir image translation for training vision tasks with challenging labels. In *Proceedings of the International Conference on Robotics and Automation (ICRA)*, pages 8291–8298. IEEE.

Yi-Lun Lee, Yi-Hsuan Tsai, Wei-Chen Chiu, and Chen-Yu Lee. 2023b. Multimodal prompting with missing modalities for visual recognition. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14943–14952.

Guangyao li, Yake Wei, Yapeng Tian, Chenliang Xu, Ji-Rong Wen, and Di Hu. 2022. Learning to answer questions in dynamic audio-visual scenarios. *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. 2024a. Videochat: Chat-centric video understanding. *Preprint*, arXiv:2305.06355.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. 2024b. Tokenpacker: Efficient visual projector for multimodal llm. *Preprint*, arXiv:2407.02392.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *Preprint*, arXiv:2311.10122.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024. Improved baselines with visual instruction tuning. *Preprint*, arXiv:2310.03744.

Mengmeng Ma, Jian Ren, Long Zhao, Davide Testuggine, and Xi Peng. 2022. Are multimodal transformers robust to missing modality? In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18177–18186.

Xiaojian Ma, Silong Yong, Zilong Zheng, Qing Li, Yitao Liang, Song-Chun Zhu, and Siyuan Huang. 2023. SQA3d: Situated question answering in 3d scenes. In *The Eleventh International Conference on Learning Representations*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2024. Video-chatgpt: Towards detailed video understanding via large vision and language models. *Preprint*, arXiv:2306.05424.

Srinivas Parthasarathy and Shiva Sundaram. 2020. Training strategies to handle missing modalities for audio-visual expression recognition. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 400–404.

Hai Pham, Paul Pu Liang, Thomas Manzini, Louis-Philippe Morency, and Barnabás Póczos. 2019. Found in translation: Learning robust joint representations by cyclic translations between modalities. In *Proceedings of the AAAI conference on artificial intelligence*, pages 6892–6899.

Yansheng Qiu, Delin Chen, Hongdou Yao, Yongchao Xu, and Zheng Wang. 2023. Scratch each other's back: Incomplete multi-modal brain tumor segmentation via category aware group self-support learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21317–21326.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2024. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *Preprint*, arXiv:2403.15388.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, Yan Lu, Jenq-Neng Hwang, and Gaoang Wang. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18221–18232.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, Hao Tian, Lewei Lu, Xizhou Zhu, Xiaogang Wang, Yu Qiao, and Jifeng Dai. 2023a. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *ArXiv*, abs/2312.09245.

Yuanzhi Wang, Yong Li, and Zhen Cui. 2023b. Incomplete multimodality-diffused emotion recognition. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. 2024. Occllama: An occupancy-language-action generative world model for autonomous driving. *ArXiv*, abs/2409.03272.

Shicai Wei, Chunbo Luo, and Yang Luo. 2023. Mmanet: Margin-aware distillation and modality-aware regularization for incomplete multimodal learning. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 20039–20049.

Zhenbang Wu, Anant Dadu, Nicholas Tustison, Brian Avants, Mike Nalls, Jimeng Sun, and Faraz Faghri. 2024. Multimodal patient representation learning with missing modalities and labels. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Johnny Xi, Jana Osea, Zuheng Xu, and Jason S Hartford. 2024. Propensity score alignment of unpaired multimodal data. *Advances in Neural Information Processing Systems*, 37:141103–141128.

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR)*.

Haofei Xu, Jing Zhang, Jianfei Cai, Hamid Rezatofighi, Fisher Yu, Dacheng Tao, and Andreas Geiger. 2023. Unifying flow, stereo and depth estimation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(11):13941–13958.

Runsen Xu, Xiaolong Wang, Tai Wang, Yilun Chen, Jiangmiao Pang, and Dahua Lin. 2024. Pointllm: Empowering large language models to understand point clouds. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 131–147. Springer.

Jiazuo Yu, Haomiao Xiong, Lu Zhang, Haiwen Diao, Yunzhi Zhuge, Lanqing Hong, Dong Wang, Huchuan Lu, You He, and Long Chen. 2024. Llms can evolve continually on modality for x-modal reasoning. *Advances in Neural Information Processing Systems*, 37:49834–49858.

Shoubin Yu, Jaehong Yoon, and Mohit Bansal. 2025. CREMA: Generalizable and efficient video-language reasoning via multimodal modular fusion. In *The Thirteenth International Conference on Learning Representations*.

Sukwon Yun, Inyoung Choi, Jie Peng, Yangfan Wu, Jingxuan Bao, Qiyiwen Zhang, Jiayi Xin, Qi Long, and Tianlong Chen. 2024. Flex-moe: Modeling arbitrary modality combination via the flexible mixture-of-experts. *Advances in Neural Information Processing Systems*, 37:98782–98805.

Changqing Zhang, Yajie Cui, Zongbo Han, Joey Tianyi Zhou, Huazhu Fu, and Qinghua Hu. 2020. Deep partial multi-view learning. *IEEE transactions on pattern analysis and machine intelligence*, 44(5):2402–2415.

Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. 2023a. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*.

Hang Zhang, Xin Li, and Lidong Bing. 2023b. Video-llama: An instruction-tuned audio-visual language model for video understanding. *Preprint*, arXiv:2306.02858.

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025. Llava-mini: Efficient image and video large multimodal models with one vision token. *Preprint*, arXiv:2501.03895.

Yunhua Zhang, Hazel Doughty, and Cees Snoek. 2023c. Learning unseen modality interaction. *Advances in Neural Information Processing Systems*, 36:54716–54726.

Mingyu Zheng, Xinwei Feng, Qingyi Si, Qiaoqiao She, Zheng Lin, Wenbin Jiang, and Weiping Wang. 2024. Multimodal table understanding. In *Proceedings of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand. Association for Computational Linguistics.

Zhuo Zhi, Ziquan Liu, Moe Elbadawi, Adam Daneshmend, Mine Orlu, Abdul Basit, Andreas Demosthenous, and Miguel Rodrigues. 2024. Borrowing treasures from neighbors: In-context learning for multimodal learning with missing modalities and data scarcity. *arXiv preprint arXiv:2403.09428*.

Tongxue Zhou, Stéphane Canu, Pierre Vera, and Su Ruan. 2021. Latent correlation representation learning for brain tumor segmentation with missing mri modalities. *IEEE Transactions on Image Processing*, 30:4263–4274.

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, Hongfa Wang, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, Caiwan Zhang, Zhifeng Li, Wei Liu, and Li Yuan. 2024. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *Proceedings of the International Conference on Learning Representations (ICLR)*. OpenReview.net.

Le Zhuo, Zewen Chi, Minghao Xu, Heyan Huang, Jianan Zhao, Heqi Zheng, Conghui He, Xian-Ling Mao, and Wentao Zhang. 2024. ProtLLM: An interleaved protein-language LLM with protein-as-word pre-training. In *Proceedings of the Association for Computational Linguistics (ACL)*, Bangkok, Thailand. Association for Computational Linguistics.

A Appendix

A.1 Efficient Flexible Multimodal Learning with Flex-M^3

To investigate the performance-computation trade-off of our generation framework, we list the parameters and the number of floating point operations (GFLOPs) per training forward of Flex-M^3 with two backbones in Table 5. From the results, we find that both compression and generation methods (Padding and Flex-M^3) incur minimal computation overhead compared to their original architectures. Especially, for BLIP-2-based architectures, with LLM frozen and PEFT techniques, we could further improve training efficiency by updating less than 1% parameters, while still benefiting from the multimodal learning performance gains.

Table 5: Comparison between Flex-M^3 and baselines on training cost on NeXT-QA with V, F, D, N modalities. p_{total} refers to total parameters (M) and $p_{train.}$ indicates all trainable parameters (M).

Modality	Avg.	p_{total}	$p_{train.}$	GFLOPs
BLIP-2	72.30	3947.65	16.83	2.47K
w/ Padding	74.22	3957.62	16.84	2.47K
w/ Flex-M^3	74.82	3966.06	25.27	2.60K
LLaVA	54.98	9307.47	9003.96	11.35K
w/ Padding	76.72	9307.47	9003.96	11.35K
w/ Flex-M^3	77.04	9307.58	9004.07	11.35K

A.2 Extra Implementation Details

The multimodal benchmarks used in this study are detailed as follows:

• **NExT-QA** (Xiao et al., 2021) is a video question answering benchmark designed to advance video understanding beyond simple descriptions towards explaining temporal actions. It focuses on causal and temporal action reasoning as well as common scene comprehension. The dataset comprises 5440 videos and approximately 52K questions. We report the results on the validation set of NExT-QA.

• **SQA3D** (Ma et al., 2023) is a compositional VideoQA task centered around situated question answering within 3D scenes. It is built upon 650 scenes from ScanNet, featuring approximately 33K diverse reasoning questions, spanning a range of capabilities, including spatial relation comprehension, commonsense understanding, navigation, and multi-hop reasoning. Following (Hong et al., 2023b), we utilize the ego-centric videos corresponding to the 3D scenes as video inputs. We report the results on the validation set.

• **MUSIC-AVQA** (li et al., 2022) is a compositional Audio-Visual Question Answering benchmark designed for comprehensive multimodal understanding and spatio-temporal reasoning over audio-visual scenes. It contains over 45K question-answer pairs derived from 9K videos. We train and evaluate baseline models and Flex-M^3 on the real video portion.

Table 6: Ablation of model modules.

Method	Avg. (%)
Baseline w/ full data	68.98
Baseline w/ full modalities	66.63
+ Generation	68.21
+ Per-modal compression	68.43
+ Cross-modal compression	**69.86**

A.3 Extra Experiment Results

We begin by directly generating multimodal encoder outputs and concatenating them as inputs for the Large Language Model (LLM). Subsequently, we integrate a per-modal compressor while keeping the concatenation, followed by incorporating a cross-modal compression mechanism. The experimental results demonstrate that the synergistic combination of these design elements achieves a Pareto-optimal balance between computational efficiency and model performance.

A.4 Comparison with Missing-Modality Baselines

Most existing missing-modality methods (Lee et al., 2023b; Zhi et al., 2024) target classification and are incompatible with the token-based MLLM paradigm. To compare against a recent representative approach, we adapt the core mechanism of Flex-MoE (Yun et al., 2024), a set of globally learnable token banks for absent modalities, optimized via an alignment loss to match real modality features, and integrate it into our BLIP-2 backbone in place of our generation module. As shown in Table 7, while the Flex-MoE bank improves over Padding by retrieving static embeddings for missing modalities, Flex-M^3 outperforms it by a substantial margin (+2.99%). This demonstrates that instance-conditioned generation, which dynamically synthesizes representations from the available anchor modalities, captures richer contextual information than a fixed global bank.

Table 7: Comparison with Flex-MoE (Yun et al., 2024) missing modality bank on NExT-QA (BLIP-2, 5K subset). **V**: Video, **F**: Flow, **D**: Depth, **N**: Norm. "Missing" indicates whether missing-modality handling is applied.

Modality	Missing	Method	Avg. (%)
V	✗	Essential	68.98
VFDN	✗	Full w/ Inc. Data	66.63
VFDN	✓	Padding	66.89
VFDN	✓	Flex-MoE Bank	68.35
VFDN	✓	Flex-M^3	**71.34**

Table 8: Effect of anchor modality availability on NExT-QA (BLIP-2, 5K subset). **V**: Video, **T**: Text, **F**: Flow, **D**: Depth, **N**: Norm. Removing the video anchor causes drastic degradation (>20%), while performance remains stable across different auxiliary modality subsets when video is present.

Modalities	Avg. (%)
VTD	70.68
VTN	70.80
VTF	70.52
VTND	71.08
VTNF	70.58
VTDF	71.08
VTDNF	70.90
TD	47.04
TN	44.64
TF	46.06

A.5 Anchor Modality Analysis

A practical design choice in Flex-M^3 is the use of text and video as consistently available anchor modalities for generation. Table 8 empirically validates this choice on a 5K NExT-QA subset. When the video anchor is present, performance remains stable across all auxiliary modality subsets (70.52–71.08%), indicating that the specific combination of auxiliary modalities has limited impact. In contrast, removing the video anchor causes a drastic accuracy drop exceeding 20% (44.64–47.04%), confirming that video carries the dominant task-relevant information, consistent with findings in prior work (Yu et al., 2025). This task-inherent dependence on video motivates our anchor design and is explicitly discussed in our Limitations section.

Table 9: Hyperparameters for training BLIP-2-based and LLaVA-based baselines and Flex-M^3.

Dataset	Modality	Batch Size / GPU	LR	Projector LR	Warmup	Epoch	Grad Accum Steps
BLIP-2							
NExT-QA	V	16	1e-4	-	1000	10	1
	V, F	16	1e-4	-	1000	10	1
	V, F, D	16	1e-4	-	1000	10	1
	V, F, D, N	8	1e-4	-	1000	10	2
SQA3D	V	16	2e-4	-	1000	20	1
	V, P, D, N	16	2e-4	-	1000	20	1
MUSIC-AVQA	V	24	2e-4	-	1000	20	1
	V, A, F, D, N	16	2e-4	-	1000	20	1
LLaVA							
NExT-QA	V	2	2e-5	1e-4	15	2	8
	V, F	2	2e-5	1e-4	15	2	8
	V, F, D	2	2e-5	1e-4	15	2	8
	V, F, D, N	2	2e-5	1e-4	15	2	8