ViPolyQwen: A Trimodal Architecture for Unified Text-Vision-Audio Embeddings via Attention Pooling and Prefix-Guided Dynamic Loss Optimization

Nguyen Anh Nguyen* (EraX) & Gtel Mobile JSC (GMobile) - Vietnam.

*Corresponding Author: nguyen@hatto.com

Abstract

(This paper is Architecture & Hypothesis. Training is still ongoing. Empirical Validation Required.)

Multimodal representation learning strives to bridge the semantic gap between disparate data types. While Vision-Language Models (VLMs) have advanced this frontier, generating unified embeddings across three modalities (text, vision, and audio) that are both versatile and computationally efficient remains largely unexplored. Existing paradigms often resort to task-specific models, separate embedding spaces, or complex multi-vector architectures. We propose `ViPolyQwen`, a trimodal approach for learning a single, high-dimensional (1024-d), unified embedding space $\mathcal{E}$ that encodes text, images, and audio. Building upon the Qwen2-VL-2B-Instruct foundation model and integrating a HuBERT audio encoder, our proposed methodology combines: (1) a heterogeneous dataset ($\mathcal{D}$, $|\mathcal{D}| > 11 \times 10^6$) encompassing six distinct multimodal interaction types—text similarity, instruction following, OCR, single/multi-turn VQA, and audio-text pairs—with emphasis on Vietnamese alongside multilingual data; (2) a **prefix-guided dynamic mixed-loss optimization strategy** that conditions the learning process, tailoring the objective function ($\mathcal{L}_{\text{NCE}}$, $\mathcal{L}_{\text{Triplet}}$, $\mathcal{L}_{\text{MSE}}$, $\mathcal{L}_{\text{Cos}}$) on a per-sample basis during training via discrete task prefixes p_i; and (3) modality-specific **Attention Pooling** mechanisms that aggregate information from each encoder's output sequence, weighting features based on learned importance. We hypothesize that this synergistic trimodal approach may yield an architecturally simpler embedding model that represents text, images, and audio in a single unified space while potentially outperforming modality-specific models on cross-modal tasks. As empirical validation is currently in progress, we present this work to stimulate discussion on unified trimodal embeddings.

1. Introduction

The proliferation of multimodal information necessitates AI systems capable of understanding and reasoning across text, vision, audio, and structured data. A cornerstone of such systems is the ability to represent diverse inputs within a shared vector space $\mathcal{E} \subset \mathbb{R}^{D_{\text{embed}}}$, enabling semantic search, cross-modal retrieval, and Retrieval-Augmented Generation (RAG) [1]. While Vision-Language Models (VLMs) [2, 3, 4] have demonstrated promising capabilities in aligning vision and language, extending this capability to three modalities—text, vision, and audio—presents several additional challenges.

Firstly, fine-tuning models typically yields embeddings specialized for a single task objective $\mathcal{L}_{\text{task}}$ (e.g., image-text contrastive loss in CLIP [2]). While effective for that specific task, these embeddings may be suboptimal for others with different geometric requirements in $\mathcal{E}$ (e.g., fine-grained text similarity regression, visual question answering, or audio-text alignment) within the *same* embedding space. This can necessitate maintaining multiple specialized models, increasing operational complexity.

Secondly, representing complex inputs often leads to multi-vector approaches [5, 6]. These methods decompose the input into multiple representations (e.g., global context $\mathbf{e}_{\text{global}}$, local patches $\{\mathbf{e}_{\text{local},i}\}$). While potentially capturing finer granularity, they introduce significant downstream complexity, requiring specialized indexing structures and multi-stage retrieval algorithms (e.g., ColBERT-style late interaction [7]) that deviate from standard, highly optimized dense vector search paradigms (like FAISS [8]).

Thirdly, the mechanism used to pool the sequence of encoder outputs into a single vector significantly impacts the

final embedding quality. Standard strategies like mean pooling ($\mathbf{c}_{\text{mean}} = \frac{1}{N}\sum \mathbf{h}_i$) may dilute salient information across modalities with different characteristics. The challenges are magnified when integrating audio, which has distinctive temporal properties and information density patterns compared to text or images.

To address these challenges, we propose `ViPolyQwen`, a unified trimodal embedding model built upon Qwen2-VL-2B-Instruct [3] for text and vision, with the addition of a HuBERT [23] model for audio processing. Our approach seeks to generate a single 1024-dimensional vector $\mathbf{e} \in \mathbb{R}^{1024}$ capable of representing diverse text, image, and audio inputs effectively. Its design is guided by three core principles:

1. **Diverse Multi-Task Trimodal Training Data:** We curate a large-scale dataset ($D = \{(x_i, y_i, \text{type}_i, ...)\}_{i=1}^{M}$, $M > 11 \times 10^6$) incorporating six distinct data formats (`type`) and associated tasks: text similarity pairs (with scores s_i), instruction-following sequences, Optical Character Recognition (OCR) / Optical Character Questioning (OCQ), single-turn Visual Question Answering (VQA), multi-turn VQA, and audio-text pairs. This diversity, with a focus on Vietnamese and substantial multilingual components, aims to foster robustness and generalization across all three modalities.
2. **Prefix-Guided Dynamic Loss Optimization:** We propose an explicit conditioning mechanism during training. Task-specific prefixes $p_i \in P = \{$`<text_pair>`, `<instr>`, `<ocr>`, `<vqa_single>`, `<vqa_multi>`, `<audio>`$\}$ are prepended to the input x_i. This prefix p_i serves as a discrete signal that dynamically selects a tailored objective function $\mathcal{L}_{\text{type}(p_i)}$ (composed of InfoNCE, Triplet Margin, MSE, Cosine Similarity components) specifically optimized for that task structure. This may allow the model, represented by parameters θ, to learn task-aware representations within the unified space $\mathcal{E}$.
3. **Modality-Specific Attention Pooling:** Departing from standard pooling, we implement separate learnable Attention Pooling mechanisms for each modality (text, image, and audio) over their respective hidden state sequences $\mathbf{H}_{\text{text}}$, $\mathbf{H}_{\text{img}}$, and $\mathbf{H}_{\text{audio}}$. This is designed to enable the model to identify and weight features based on learned importance (α_i weights for each hidden state $\mathbf{h}_i$), producing more contextually relevant intermediate representations before projection to the final unified embedding $\mathbf{e}$.

We hypothesize and aim to validate through ongoing work that the combination of diverse multi-task learning across three modalities, prefix-guided dynamic loss adaptation, and attention-based feature aggregation might enable `ViPolyQwen` to produce unified 1D embeddings that balance performance with architectural simplicity. This work has been conducted in collaboration with the AI technology team at Gtel Mobile JSC (GMobile), whose support has been valuable in this research endeavor.

2. Related Work

Our work builds upon and relates to several research directions:

- **Multimodal Contrastive Learning (e.g., CLIP, ALIGN):** Foundational models like CLIP [2] and ALIGN [9] have demonstrated effective image-text alignment through contrastive learning across large datasets. However, a single contrastive objective, while effective for retrieval, may not optimally capture the nuances required for diverse downstream tasks within the *same* embedding space. Our proposed `ViPolyQwen` approach extends beyond bimodal setups to incorporate audio alongside text and images, applying multiple loss formulations within a single training framework, guided by task type.
- **Sentence & Text Embeddings (e.g., Sentence-BERT):** Fine-tuning approaches like Sentence-BERT [10] typically focus on optimizing for a specific pair-based task structure. Applying such a focused approach naively to trimodal, multi-task data might create embeddings biased towards one structure or modality. The dynamic loss selection mechanism in our proposed approach aims to apply appropriate optimization for each data type encountered, balancing the requirements of text, image, and audio modalities.
- **Audio-Text Alignment Models:** Recent work in audio-text alignment has shown promising results in creating joint representations for speech transcription, audio captioning, and audio retrieval tasks [24, 25]. However, these models typically create specialized embeddings optimized for audio-text pairs specifically, rather than unified representations across three modalities. Our approach incorporates audio-text alignment within a broader multimodal framework.
- **Document AI & Multi-Vector Representations (e.g., ColPali):** Addressing the complexity of structured documents, multi-vector approaches like ColPali [5] dedicate separate representations for different granularities. Our prefix-guided approach, coupled with modality-specific Attention Pooling, explores an alternative possibility: whether a *single* vector could effectively encode task-relevant nuances and salient

features to handle diverse tasks across text, image, and audio modalities, thereby maintaining architectural simplicity.

- **Pooling Mechanisms:** While mean/max/last-token pooling are computationally efficient, they may not optimally aggregate information across different modalities with distinct characteristics. Our modality-specific Attention Pooling mechanisms attempt to balance effectiveness and efficiency through learnable context vector approaches tailored to each modality's structure.
- **Multi-Task Learning & Dynamic Loss:** Training models on multiple tasks simultaneously can improve generalization [12]. Dynamically selecting or weighting losses may help navigate conflicting gradient signals [13, 14]. Our prefix-guided mechanism provides an *explicit, discrete* signal for selecting task-optimized loss combinations, potentially ensuring appropriate geometric constraints are applied during optimization for each sample type across all three modalities.
- **Vietnamese & Cross-Lingual Models:** Our work addresses the need for multimodal embeddings for Vietnamese, leveraging substantial native data alongside multilingual resources to potentially foster both in-language performance and cross-lingual capabilities [15].

The proposed contribution of `ViPolyQwen` lies in the integration of: (1) a powerful VLM backbone augmented with an audio encoder, (2) conditioning the learning process on diverse task structures via prefix signals coupled with dynamic loss selection, and (3) employing modality-specific Attention Pooling to generate a unified embedding. This approach seeks to address limitations of bimodal models, single-objective training, task-specific fine-tuning, and multi-vector representation architectures.

3. Methodology

3.1 Model Architecture

The `ViPolyQwen` embedder builds upon the `Qwen/Qwen2-VL-2B-Instruct` model [3] for text and image processing, augmented with a HuBERT model [23] for audio processing. The core components involved in generating the final 1D embedding $\mathbf{e} \in \mathbb{R}^{1024}$ are:

1. **Trimodal Encoders:**
   - **Qwen-VL Processor & Encoder:** Text and image inputs are processed and tokenized by the `AutoProcessor`. During training, textual inputs are augmented with task prefixes p_i (Section 3.4). The multimodal encoder processes these inputs, yielding a sequence of final layer hidden states:

$$\mathbf{H}_{\text{text/img}} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N] \in \mathbb{R}^{N \times D_{\text{hidden}}}$$

   where $\mathbf{h}_i$ represents the contextualized state for the i-th token or visual patch, and $D_{\text{hidden}} = 2048$ for Qwen2-VL-2B.

   - **HuBERT Audio Encoder:** Audio inputs are processed by the HuBERT encoder [23], producing a sequence of hidden states:

$$\mathbf{H}_{\text{audio}} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_K] \in \mathbb{R}^{K \times D_{\text{audio}}}$$

   where $\mathbf{h}_i$ represents the audio feature at time step i, and $D_{\text{audio}} = 768$ for HuBERT.

2. **Modality-Specific Attention Pooling Layers:** Each modality employs a dedicated Attention Pooling layer (Section 3.2) to aggregate its respective hidden state sequence into a single context vector:

- $\mathbf{c}_{\text{text/img}} \in \mathbb{R}^{D_{\text{hidden}}}$ for text/image from Qwen-VL
- $\mathbf{c}_{\text{audio}} \in \mathbb{R}^{D_{\text{audio}}}$ for audio from HuBERT

3. **Modality-Specific Projection Heads:** Each modality has a dedicated trainable projection head that maps the pooled context vector to the target embedding dimension $D_{\text{embed}} = 1024$. Each consists of a linear transformation followed by Layer Normalization [16]:

$$\mathbf{p}_{\text{text/img}} = \text{LayerNorm}(\mathbf{W}_{\text{proj-text/img}}\mathbf{c}_{\text{text/img}})$$
$$\mathbf{p}_{\text{audio}} = \text{LayerNorm}(\mathbf{W}_{\text{proj-audio}}\mathbf{c}_{\text{audio}})$$

where $\mathbf{W}_{\text{proj}-\text{text/img}} \in \mathbb{R}^{D_{\text{embed}} \times D_{\text{hidden}}}$ and $\mathbf{W}_{\text{proj}-\text{audio}} \in \mathbb{R}^{D_{\text{embed}} \times D_{\text{audio}}}$ are the learnable weight matrices of the linear layers (biases are omitted).

4. **L2 Normalization:** The final embedding $\mathbf{e} \in \mathbb{R}^{D_{\text{embed}}}$ is obtained by L2 normalizing the appropriate projected vector based on the input modality:

$$\mathbf{e} = \frac{\mathbf{p}_{\text{modality}}}{||\mathbf{p}_{\text{modality}}||_2}$$

This ensures all embeddings reside on the unit hypersphere, facilitating cosine similarity comparisons across modalities.

3.2 Modality-Specific Attention Pooling Mechanisms

To derive the context vector from each encoder's hidden state sequence, we implement separate Attention Pooling mechanisms for each modality. Unlike mean pooling or last-token pooling, Attention Pooling computes a weighted average where weights reflect the learned importance of each hidden state.

For each modality, the process follows these steps:

1. **Learnable Context Vector:** Each modality has a dedicated trainable parameter vector (denoted `attention_context_vector`), initialized randomly and updated during training:
   - $\mathbf{v}_{\text{text/img}} \in \mathbb{R}^{D_{\text{hidden}}}$ for text/image
   - $\mathbf{v}_{\text{audio}} \in \mathbb{R}^{D_{\text{audio}}}$ for audio

   These vectors function as learnable "queries" representing the concept of "salience" within each modality's context.

2. **Attention Scores:** Unnormalized attention scores are computed for each hidden state via dot product:

$$u_i^{\text{text/img}} = \mathbf{h}_i^T \mathbf{v}_{\text{text/img}}$$
$$u_i^{\text{audio}} = \mathbf{h}_i^T \mathbf{v}_{\text{audio}}$$

3. **Masking:** Scores corresponding to padded positions (identified via the attention mask $\mathbf{M} \in \{0,1\}^N$) are masked:

$$u'_i = \begin{cases} u_i & \text{if } M_i = 1 \\ -\infty & \text{if } M_i = 0 \end{cases}$$

4. **Attention Weights:** The masked scores are normalized using softmax for each modality:

$$\alpha_i = \frac{\exp(u'_i)}{\sum_j \exp(u'_j)}$$

5. **Weighted Average:** The final pooled context vector for each modality is computed:

$$\mathbf{c}_{\text{text/img}} = \sum_i \alpha_i \mathbf{h}_i$$
$$\mathbf{c}_{\text{audio}} = \sum_i \alpha_i \mathbf{h}_i$$

This modality-specific mechanism allows the model to focus on potentially informative parts of each sequence (e.g., keywords in text, salient visual regions in images, distinctive audio patterns) when constructing the 1D representation.

3.3. Input Processing and Modality Detection

The model includes a sophisticated input processing pipeline that can handle various input combinations:

1. **Text-Only Inputs:** Processed directly through the Qwen-VL text encoder path.
2. **Image+Text Inputs:** Processed through the Qwen-VL multimodal encoder.
3. **Audio-Only Inputs:** Processed through the HuBERT encoder, with a special `<audio>` token prepended to an empty text sequence.
4. **Audio+Text Inputs:** The audio is processed through HuBERT while the text (often a transcript) is prepended with the `<audio>` token and processed through the Qwen-VL text encoder.

The model determines the appropriate encoding pathway through explicit flags: - `has_image` signals the presence of visual content - `has_audio` signals the presence of audio content

This allows flexible handling of unimodal, bimodal, or trimodal inputs within the same framework.

3.4 Prefix-Guided Input Representation & Conditioning (Training)

During training, the `MixedBatchCollator` preprocesses each sample $(x_i, y_i, \text{type}_i, ...)$. Based on `data_type`, a prefix $p_i \in P = \{\texttt{<text\_pair>}, \texttt{<instr>}, \texttt{<ocr>}, \texttt{<vqa\_single>}, \texttt{<vqa\_multi>}, \texttt{<audio>}\}$ is prepended to the textual input x_i, yielding $x_i' = (\text{prefix}(p_i), x_i)$.

This explicit prefix p_i acts as a **conditioning signal**. Let the embedding function be $f_\theta : (X', P) \mapsto \mathcal{E}$. The prefix p_i directly influences the selection of the loss function $\mathcal{L}_{\text{type}(p_i)}$ (Section 4.2). The gradient contributing to the update of shared parameters θ is thus task-dependent:

$$\nabla_\theta \mathcal{L}_{\text{batch}} = \frac{1}{B} \sum_{i=1}^{B} \nabla_\theta \mathcal{L}_{\text{type}(p_i)}(f_\theta(x_i'), f_\theta(y_i'))$$

This explicit conditioning is hypothesized to enable task specialization *within* the unified space $\mathcal{E}$. For inference on general data, no prefix is used ($p = \text{None}$), yielding a general-purpose embedding $f_\theta(x, \text{None})$.

3.4.1 Theoretical Foundations for Prefix-Guided Conditioning

The necessity of prefix tokens in the ViPolyQwen architecture emerges from fundamental challenges in creating unified trimodal embedding spaces. We identify and address three core theoretical issues that motivate our approach:

Task Ambiguity and Input Space Entanglement

For heterogeneous multimodal data types that share structural similarities, the model may face inherent ambiguity in determining the appropriate embedding strategy. Formally, we can express this as an input classification problem $\mathcal{C} : \mathcal{X} \to \mathcal{T}$ that maps inputs to their appropriate task types, where $\mathcal{T} = \{\text{text\_pair}, \text{ocr}, \text{instr}, \text{vqa\_single}, \text{vqa\_multi}, \text{audio}\}$. Without explicit signals, the function $\mathcal{C}$ becomes ill-defined due to overlapping input distributions:

$$P(\mathcal{X}_{\text{ocr}}) \cap P(\mathcal{X}_{\text{vqa\_single}}) \neq \emptyset$$
$$P(\mathcal{X}_{\text{text\_pair}}) \cap P(\mathcal{X}_{\text{instr}}) \neq \emptyset$$
$$P(\mathcal{X}_{\text{audio}}) \cap P(\mathcal{X}_{\text{ocr}}) \neq \emptyset \text{ (e.g., speech recognition vs. text-in-image)}$$

For instance, an image with text and a question could represent either an OCR task (requiring precise text localization) or a visual question-answering task (requiring broader scene understanding). Similarly, an audio clip with spoken text could be treated as pure audio content for embedding or as a transcription task. Without additional signaling, the model must implicitly infer task type, potentially introducing noise into the learning process. Prefix tokens provide an explicit, unambiguous signal p_i that resolves this classification uncertainty, formally:

$$P(\mathcal{T} = t | \mathcal{X} = x, P = p_t) = 1$$

where p_t is the task-specific prefix for task $t \in \mathcal{T}$.

Conflicting Geometric Constraints in Embedding Space

Each task-specific loss function imposes distinct geometric constraints on the embedding space $\mathcal{E}$. We can formalize these constraints as manifolds or regions within $\mathcal{E}$ where:

- For InfoNCE loss ($\mathcal{L}_{\text{NCE}}$): Positive pairs should be closer than all negatives by a certain margin in a batch-dependent context.
- For Triplet Margin loss ($\mathcal{L}_{\text{Triplet}}$): Positive pairs should maintain a fixed minimum distance from the hardest negative.
- For MSE Similarity Regression ($\mathcal{L}_{\text{MSE}}$): Embedding similarity should match a continuous target score, creating a regression manifold.
- For Cosine Similarity Maximization ($\mathcal{L}_{\text{Cos}}$): Directly maximizes alignment between specific pairs.

These constraints can conflict when applied simultaneously to inputs from different tasks or modalities. Formally, we can express the optimal embedding regions for different losses as:

$$\mathcal{E}^{\mathcal{L}_{\text{NCE}}}_{\text{optimal}} \cap \mathcal{E}^{\mathcal{L}_{\text{MSE}}}_{\text{optimal}} \neq \mathcal{E}^{\mathcal{L}_{\text{NCE}}}_{\text{optimal}} \text{ and } \neq \mathcal{E}^{\mathcal{L}_{\text{MSE}}}_{\text{optimal}}$$

Our prefix-guided approach addresses this by providing task context that supports "multimodal loss disambiguation":

$$\nabla_\theta \mathcal{L}(f_\theta(x'_i), f_\theta(y'_i)) = \nabla_\theta \mathcal{L}_{\text{type}(p_i)}(f_\theta(x'_i), f_\theta(y'_i))$$

This allows the model to navigate the trade-offs between competing geometric constraints in a principled manner, activating appropriate optimization pressures for each sample based on its task characteristics and modality combination.

Neuron Activation Specialization and Knowledge Transfer

Prefix tokens enable what we term "conditional activation patterns" within the model's parameters. With a conditional input p_i, certain neurons or attention heads may specialize in task-specific features while maintaining shared representations, formally:

$$\mathbf{h}^{(l)}_j = \sigma\left(\mathbf{W}^{(l)}_j \mathbf{h}^{(l-1)} + \mathbf{b}^{(l)}_j\right) \cdot g(p_i, \mathbf{h}^{(l-1)})$$

where $g(p_i, \mathbf{h}^{(l-1)})$ represents a modulation function influenced by the prefix token. For example, when processing an OCR sample (prefix `<ocr>`), neurons specialized in text localization may exhibit higher activation levels, while for audio samples, neurons attuned to temporal patterns might dominate.

This controlled form of specialization offers two key benefits:

1. **Parameter Efficiency**: Rather than training entirely separate models for each task and modality, parameters are shared with targeted conditional activations.
2. **Cross-Task Knowledge Transfer**: Learning from one task implicitly benefits others through shared parameters, while task-specific aspects remain differentiated via the prefix conditioning signal.

The interplay between modality-specific Attention Pooling and prefix-guided conditioning creates a synergistic effect: Attention Pooling focuses on extracting contextually important features from each modality's sequence, while prefix tokens guide which features should be considered important in the current task context.

3.4.2 Prefix Usage in Training vs. Inference

A distinguishing aspect of our approach is the asymmetry between training and inference prefix usage:

- **During Training**: Every input explicitly includes a task-specific prefix to guide loss selection and facilitate task-aware representation learning.

- **During Inference (General Case)**: For most common embedding scenarios (text chunks, single images, audio clips), no prefix is required. The model learns to produce generalized embeddings that capture unified multimodal understanding.
- **During Inference (Specialized Case)**: For specific task scenarios like OCR querying, focused VQA retrieval, or audio-text alignment, prefixes can optionally be included to "steer" the embedding toward task-optimized regions of the embedding space.

This design offers a unique compromise: encoding task-specialized knowledge during training while providing simplified, prefix-free inference for general use cases. When fine-grained control or specialized capabilities are needed, prefixes can be selectively reintroduced at inference time.

The general-purpose embedding function without prefixes is defined as:

$$\mathbf{e}_{\text{general}}(x) = f_\theta(x, \text{None})$$

While the task-steered embedding function with prefixes is:

$$\mathbf{e}_{\text{task}}(x, t) = f_\theta(x, p_t)$$

This dual interface balances simplicity for common use cases with the power of task-specific optimization when required.

4. Training Paradigm

4.1 Dataset Composition

The model is trained on a composite dataset $\mathcal{D}$ (>11M samples) covering:

- **Text Similarity (`<text_pair>`):** Text pairs (x_i, y_i) with similarity scores s_i. (Vi/En/Zh)
- **Instruction Following (`<instr>`):** (Instruction, Output) pairs (x_i, y_i).
- **OCR/OCQ (`<ocr>`):** (Image(s)+Question, Answer) triples (x_i, y_i).
- **Single/Multi-turn VQA (`<vqa_...>`)**: (Image(s)+Context/Question, Answer) triples (x_i, y_i).
- **Audio-Text Pairs (`<audio>`)**: (Audio clip, Transcript) pairs (x_i, y_i).

The dataset comprises predominantly Vietnamese (approximately 60%), with English (approximately 30%) and Chinese (approximately 10%) portions.

4.2 Prefix-Guided Dynamic Mixed-Loss Optimization

The training objective dynamically applies task-specific losses based on prefix p_i. Let $(\mathbf{e}_{a,i}, \mathbf{e}_{b,i}) = (f_\theta(x'_i), f_\theta(y'_i))$ be normalized embeddings.

- **For p_i = `<text_pair>`:** Combines contrastive loss and score regression.

$$\mathcal{L}_{\text{text\_pair}} = \lambda_{\text{nce}}\mathcal{L}_{\text{NCE}}(\mathbf{e}_{a,i}, \mathbf{e}_{b,i}, \mathcal{B}, T) + \lambda_{\text{mse}}\mathcal{L}_{\text{MSE}}(\mathbf{e}_{a,i}, \mathbf{e}_{b,i}, s_i)$$

  where $T = 0.07$, $\lambda_{\text{nce}} = \lambda_{\text{mse}} = 1.0$, $\mathcal{L}_{\text{MSE}} = (\frac{1}{2}(\mathbf{e}_{a,i}^T\mathbf{e}_{b,i} + 1) - s_i)^2$, and $\mathcal{L}_{\text{NCE}}$ is symmetric InfoNCE over batch $\mathcal{B}$:

$$\mathcal{L}_{\text{NCE}} = -\frac{1}{2B}\sum_{k=1}^{B}\left[\log\frac{\exp(S_{k,k}/T)}{\sum_{j=1}^{B}\exp(S_{k,j}/T)} + \log\frac{\exp(S_{k,k}/T)}{\sum_{j=1}^{B}\exp(S_{j,k}/T)}\right]$$

  with $S_{kj} = \mathbf{e}_{a,k}^T\mathbf{e}_{b,j}$.

- **For p_i = `<instr>`:** Combines contrastive loss and direct similarity maximization.

$$\mathcal{L}_{\text{instr}} = \lambda_{\text{nce}}\mathcal{L}_{\text{NCE}}(\mathbf{e}_{a,i}, \mathbf{e}_{b,i}, \mathcal{B}, T) + \lambda_{\text{cos}}\mathcal{L}_{\text{Cos}}(\mathbf{e}_{a,i}, \mathbf{e}_{b,i})$$

where $\lambda_{\text{cos}} = 1.0$ and $\mathcal{L}_{\text{Cos}} = (1 - \mathbf{e}_{a,i}^T \mathbf{e}_{b,i})$.

- **For $p_i \in \{$`<ocr>`, `<vqa_single>`, `<vqa_multi>`$\}$:** Combines contrastive loss and triplet margin loss.

$$\mathcal{L}_{\text{ocr/vqa}} = \lambda_{\text{nce}} \mathcal{L}_{\text{NCE}}(\mathbf{e}_{a,i}, \mathbf{e}_{b,i}, \mathcal{B}, T) + \lambda_{\text{trip}} \mathcal{L}_{\text{Triplet}}(\mathbf{e}_{a,i}, \mathbf{e}_{b,i}, \mathcal{N}_i, m', T)$$

where $\lambda_{\text{trip}} = 1.0$ (or 1.5 for multi-turn), $m' = 0.2$ (or 0.3 for multi-turn), $\mathcal{N}_i = \{\mathbf{e}_{b,j} \mid j \neq i\}$, and

$$\mathcal{L}_{\text{Triplet}} = \max\left(0, \max_{\mathbf{e}_n \in \mathcal{N}_i} \frac{\mathbf{e}_{a,i}^T \mathbf{e}_n}{T} - \frac{\mathbf{e}_{a,i}^T \mathbf{e}_{b,i}}{T} + m'\right)$$

- **For $p_i =$ `<audio>`:** Combines contrastive loss, direct similarity maximization, and triplet margin loss.

$$\mathcal{L}_{\text{audio}} = \lambda_{\text{nce}} \mathcal{L}_{\text{NCE}}(\mathbf{e}_{a,i}, \mathbf{e}_{b,i}, \mathcal{B}, T) + \lambda_{\text{cos}} \mathcal{L}_{\text{Cos}}(\mathbf{e}_{a,i}, \mathbf{e}_{b,i}) + \lambda_{\text{trip}} \mathcal{L}_{\text{Triplet}}(\mathbf{e}_{a,i}, \mathbf{e}_{b,i}, \mathcal{N}_i, m'', T)$$

where $\lambda_{\text{cos}} = 1.0$, $\lambda_{\text{trip}} = 1.0$, and $m'' = 0.2$, providing a comprehensive optimization approach for audio-text alignment.

The overall batch loss is $\mathcal{L}_{\text{batch}} = \frac{1}{B} \sum_{i=1}^{B} \mathcal{L}_{\text{type}(p_i)}$.

4.3 Implementation Details (ongoing):

- **Hardware:** 4x NVIDIA H100 GPUs (94GB VRAM).
- **Framework:** Hugging Face `accelerate` with FSDP ZeRO-3.
- **Precision:** `bfloat16` mixed precision, Flash Attention 2.
- **Optimizer:** AdamW [17].
- **Learning Rate:** 1×10^{-5} initial 5% warmup, with subsequent cosine decay
- **Batch Size:** Per-device 24, gradient accumulation 8 (Global: 768).
- **Sequence Length:** 8192 tokens.
- **Training Duration:** 2-3 epochs (approximately 15-24 days).
- **Regularization:** Weight decay 0.001, max gradient norm 1.0.
- **Loss Parameters:** $T = 0.07$, $m = 0.2$ (base). λ's = 1.0.
- **Tokenizer:** Extended Qwen-VL tokenizer with new prefix tokens (including `<audio>`) and embedding model's layer resized.

4.4 Multimodal Data Collation and Batching

A critical component of our trimodal architecture is the `MixedBatchCollator`, which handles the complex process of combining different data types and modalities into coherent batches:

1. **Modality Detection:** The collator identifies which modalities are present in each example (text, image, audio) and sets appropriate flags.
2. **Audio Processing:** For audio examples, the collator extracts audio files, processes them with the HuBERT processor to obtain audio features, and creates appropriate attention masks.
3. **Image Processing:** For visual examples, images are loaded and processed using the Qwen-VL processor.
4. **Prefix Insertion:** Based on the data type, appropriate task prefixes are prepended to text inputs.
5. **Batch Creation:** The collator assembles batches containing the following components:
    - `input_ids_a`, `attention_mask_a`: Text tokens for the first part of pairs
    - `input_ids_b`, `attention_mask_b`: Text tokens for the second part of pairs
    - `pixel_values_a`, `pixel_values_b`: Visual features when images are present
    - `audio_features_a`, `audio_attention_mask_a`: Audio features and masks
    - `audio_features_b`, `audio_attention_mask_b`: Audio features for the paired side
    - `has_audio_a`, `has_audio_b`: Boolean flags indicating audio presence
    - `has_image_a`, `has_image_b`: Boolean flags indicating image presence
    - `data_type`: String identifiers of example types
    - `similarity_scores`: Numeric scores for contrastive pairs (when available)

This collation process ensures that the model can seamlessly handle heterogeneous data types while enforcing appropriate batch structure for the prefix-guided loss functions.

4.5 Complementary Roles of Attention Pooling and Prefix-Guided Conditioning

While both modality-specific Attention Pooling and prefix-guided conditioning contribute to improved representations, they serve distinct functional roles within the model architecture. Understanding their synergistic relationship clarifies the overall design philosophy:

Modality-Specific Attention Pooling focuses on the problem of information extraction from each encoder's output sequence. It addresses the question: "How do we best summarize the sequence of hidden states into a single vector for each modality?" By learning to assign attention weights α_i to each token/patch/audio representation $\mathbf{h}_i$, it creates a nuanced weighted average that emphasizes the most salient features for the final embedding.

In contrast, **Prefix-Guided Conditioning** addresses the problem of task disambiguation and appropriate loss application. It answers the question: "Which optimization constraints should be prioritized for this particular input across modalities?" The prefix tokens signal to the model which task family the current input belongs to, enabling the application of the most appropriate loss function combination.

The relationship between these mechanisms can be expressed formally as:

1. **Modality-Specific Attention Pooling** transforms the encoder hidden states into context vectors:

$$\mathbf{c}_{\text{modality}} = \sum_i \alpha_i \mathbf{h}_i$$

2. **Prefix-Guided Conditioning** influences which loss function is applied to these context vectors after projection:

$$\mathcal{L} = \mathcal{L}_{\text{type}(p_i)}(f_\theta(x_i'), f_\theta(y_i'))$$

This complementary design enables a form of "dual adaptation" - Attention Pooling adapts the feature extraction process to each input's content and modality characteristics, while prefix conditioning adapts the optimization process to each input's task structure.

5. Experimental Design and Evaluation Plan

As `ViPolyQwen` is currently undergoing training, we outline a comprehensive evaluation plan designed to assess its capabilities and validate our core hypotheses upon completion.

5.1 Target Benchmarks and Metrics

Our evaluation strategy encompasses standard cross-modal benchmarks, trimodal tasks, tasks specific to Vietnamese, and assessments relevant to document understanding:

- **Image-Text Retrieval (Zero-Shot):** Evaluation on established datasets like MS-COCO 5k Captions [18] and Flickr30k [19]. Standard metrics including Recall@K (R@1, R@5, R@10) and Mean Rank (MeanR) will be computed for both Text-to-Image (T->I) and Image-to-Text (I->T) directions.
- **Audio-Text Retrieval:** Evaluation on audio-text retrieval benchmarks such as Clotho [26] and AudioCaps [27]. Metrics include Recall@K for both Audio-to-Text (A->T) and Text-to-Audio (T->A) directions.
- **Cross-Modal Search (Trimodal):** A novel evaluation task where queries from one modality (text, image, or audio) are used to retrieve relevant content from the other two modalities. This will assess the model's ability to establish coherent semantic relationships across all three modalities within the same embedding space.
- **Vietnamese Semantic Textual Similarity (STS):** Performance will be measured on the ViSTS subset of the ViTextEval suite [20], using Spearman's rank correlation coefficient (ρ) between the cosine similarity of generated embeddings and human judgments.

- **Document Context Retrieval (Proxy for Document VQA):** Using datasets like DocVQA [21], we will assess the ability of embeddings to retrieve document pages containing answers to visual questions. Metrics will include Page Retrieval Accuracy@K (Acc@1, Acc@5), serving as a proxy for the embedding's utility in supporting document understanding tasks.
- **Ablation Studies:** A held-out internal validation set (5k samples) will be used to quantify the individual contributions of key components:
  - Attention Pooling vs. Mean Pooling for each modality
  - Dynamic Loss vs. Single Objective
  - Trimodal vs. Bimodal training approaches

5.2 Baselines for Comparison

To contextualize the performance of our approach, we plan to compare against several relevant baselines:

- **Strong Image-Text Models:** CLIP (ViT-L/14) [2] as a foundational contrastive learning baseline for image-text tasks.
- **Audio-Text Models:** HuBERT [23] and AudioCLIP [28] as baselines for audio-text alignment tasks.
- **Base Models with Simplified Pooling:**
  - The Qwen2-VL-2B-Instruct model [3] with standard mean pooling for text-image tasks
  - HuBERT with mean pooling for audio tasks
- **Multimodal and Multilingual Models:** Representative multilingual text-image models (e.g., mCLIP adaptations [22]) for cross-lingual evaluation.
- **Ablation Variants:**
  - `ViPolyQwen-MeanPool`: Our model trained with the full prefix-guided dynamic loss suite but utilizing mean pooling instead of Attention Pooling.
  - `ViPolyQwen-NCEOnly`: Our model trained with Attention Pooling but employing only the InfoNCE loss component for all data types.
  - `ViPolyQwen-NoAudio`: A bimodal variant without the audio encoder, to assess the impact of adding the third modality.
- **Conceptual Comparison:** We will qualitatively discuss architectural trade-offs and potential performance implications relative to multi-vector paradigms like ColPali [5], particularly concerning system complexity and deployment efficiency.

5.3 Evaluating the Specific Contributions of the Trimodal Architecture

To rigorously assess the impact of our trimodal approach, we will include additional experimental conditions focused specifically on the integration of audio alongside text and images:

- **Modality Isolation Tests:** Evaluating performance when embedding each modality independently versus the unified approach.
- **Audio Representation Quality:** Comparing the quality of audio embeddings when trained in isolation versus within our trimodal framework.
- **Cross-Modal Transfer:** Measuring if training on all three modalities improves performance on bimodal tasks (e.g., does audio-text training improve image-text performance?).
- **Modality Ablation Studies:** Selective removal of specific modalities during training to quantify their contribution to overall performance.

5.4 Evaluating the Specific Contributions of Prefix-Guided Conditioning

To rigorously assess the impact of our prefix-guided approach, we will include additional experimental conditions focused specifically on this mechanism:

- **Implicit Task Inference**: A model variant trained without explicit prefixes that must infer task type from input structure alone.
- **Fixed Loss Weighting**: A model using the same loss combination (weighted sum of all loss components) for all samples, regardless of task type.

- **Prefix Ablation by Task**: Selective removal of specific prefixes to measure their impact on corresponding task performance.

For each variant, we will evaluate both task-specific performance (e.g., OCR accuracy, similarity regression, audio transcription) and cross-task generalization to quantify how prefix-guided dynamic loss optimization contributes to the model's capabilities.

6. Research Hypotheses

This research explores several hypotheses regarding our proposed methodology. The ongoing training and subsequent evaluation are designed to examine these propositions. We present them to invite discussion from the research community:

1. **H1: On the Effectiveness of Modality-Specific Attention Pooling:** We hypothesize that the learnable Attention Pooling mechanism tailored to each modality may capture more salient information compared to standard pooling approaches. By dynamically weighting features based on learned importance, it might produce more discriminative 1D embeddings, particularly for complex inputs like documents containing text, audio with varying acoustic properties, or complex visual scenes.
2. **H2: On Prefix-Guided Dynamic Loss and Trimodal Task Versatility:** We propose that explicitly conditioning the training on task type via prefixes and applying tailored loss functions may be beneficial for achieving robust performance across diverse tasks and modality combinations in our training data. A single contrastive objective might be suboptimal compared to the dynamic loss strategy, which applies task-specific geometric constraints within the unified embedding space for text, image, and audio inputs.
3. **H3: On the Viability of Unified Single-Vector Trimodal Representation:** We explore whether the combination of powerful foundational models (Qwen-VL and HuBERT), diverse multi-task dynamic training, and modality-specific Attention Pooling might enable encoding sufficient trimodal nuance within a single vector to be competitive with more complex architectures, while providing deployment advantages (standard indexing/search infrastructure, potentially lower latency).
4. **H4: On Multilingual and Vietnamese Performance:** Given the substantial proportion of Vietnamese data in our training set, we aim to investigate whether our trimodal approach can establish a viable baseline for Vietnamese multimodal embedding tasks, performing competitively with models specifically optimized for the language across all three modalities.
5. **H5: On Cross-Modal Transfer Benefits:** We hypothesize that training on a trimodal dataset enables beneficial knowledge transfer between modalities, potentially improving performance even on bimodal tasks compared to models trained only on those modalities. For example, audio-text alignment might enhance the model's capabilities for processing speech in videos or spoken instructions.
6. **H6: On the Necessity of Explicit Task Disambiguation:** We hypothesize that the explicit prefix-guided approach will outperform implicit task inference, particularly for structurally similar inputs with different semantic requirements (e.g., audio recognition vs. audio-text alignment, or OCR vs. general VQA on text-containing images). This hypothesis addresses whether the benefits of explicit conditioning outweigh the simplicity of prefix-free training.
7. **H7: On Conflicting Geometric Constraints in Trimodal Space:** We propose that different task types across the three modalities inherently benefit from different loss functions due to their distinct geometric requirements in embedding space. The dynamic loss selection mechanism should demonstrate measurable advantages over applying any single loss function across all tasks, or even over applying a fixed weighted combination of losses.

Call for Discussion: As the training process for such a large-scale trimodal model requires significant resources, we present these hypotheses and our experimental design prior to obtaining final results to invite feedback from the community. We welcome suggestions for additional benchmarks, baselines, or insights regarding our proposed approach.

7. Conclusion and Future Directions

In this paper, we have introduced `ViPolyQwen`, a framework for learning unified trimodal embeddings within a single vector space. The approach integrates three key components: (1) a diverse multi-task training dataset spanning text, image, and audio modalities; (2) a prefix-guided mechanism for dynamically selecting task-optimized

loss functions; and (3) modality-specific Attention Pooling layers for feature aggregation. The central hypothesis is that this integration might yield embeddings that are versatile across different modalities and tasks while maintaining architectural simplicity.

The immediate next step is completing the ongoing training phase, followed by rigorous empirical validation through the evaluation plan outlined in Section 5. This will involve comparing our approach against established baselines and conducting ablation studies to understand the contribution of each component. Upon completion of this validation, we plan to release model checkpoints, evaluation code, and usage guidelines to facilitate further research.

Future Research Directions: Subject to empirical validation of our approach, several promising research directions may be explored:

- **Scaling Effects:** Investigating how the proposed methodology performs when applied to larger foundation models.
- **Modality Expansion:** Exploring the potential integration of additional modalities (e.g., video, 3D data) into the unified embedding space using similar principles.
- **Application Studies:** Examining the practical benefits of the proposed embeddings in downstream applications such as multimodal retrieval systems, cross-modal search, and document understanding platforms.
- **Architectural Refinements:** Further research into attention mechanisms and loss formulations to enhance representation quality across modalities.
- **Adaptive Prefix Inference:** Developing mechanisms that could automatically infer the most appropriate prefix during inference time based on input characteristics, potentially offering task-optimized embeddings without requiring explicit user specification.
- **Theoretical Analysis:** Deeper mathematical analysis of how different loss functions shape the embedding space geometry in a trimodal context and how prefix-guided conditioning mediates between potentially conflicting geometric constraints.
- **Multimodal Fusion Studies:** Investigating different strategies for fusing information from the three modalities, potentially introducing cross-modal attention mechanisms to enhance the integration of information.

We hope that the principles and methodologies proposed in this work contribute to the ongoing conversation about efficient, versatile multimodal representations, particularly for complex inputs that span text, image, and audio modalities. By integrating these three distinct data types within a unified embedding framework, ViPolyQwen represents a step toward more comprehensive and flexible AI systems capable of processing the rich, multimodal world we inhabit.

References

[1] P. Lewis, E. Perez, A. Piktus, et al., "Retrieval-augmented generation for knowledge-intensive NLP tasks," in Advances in Neural Information Processing Systems (NeurIPS), 2020.

[2] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning transferable visual models from natural language supervision," in International Conference on Machine Learning (ICML), 2021.

[3] J. Bai, S. Bai, S. Yang, et al., "Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond," arXiv preprint arXiv:2308.12966, 2023.

[4] J.-B. Alayrac, J. Donahue, P. Dieleman, et al., "Flamingo: a visual language model for few-shot learning," in Advances in Neural Information Processing Systems (NeurIPS), 2022.

[5] M. Faysse, H. Sibille, T. Wu, et al., "Colpali: Efficient document retrieval with vision language models," arXiv preprint arXiv:2407.01449, 2024.

[6] Z. Zhang, R. Müller, W. Morris, et al., "Beyond pixels and patches: Utilizing vlm for document information extraction," arXiv preprint arXiv:2310.00425, 2023.

[7] O. Khattab and M. Zaharia, "Colbert: Efficient and effective passage search via contextualized late interaction over BERT," in Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR), 2020.

[8] J. Johnson, M. Douze, and H. Jégou, "Billion-scale similarity search with GPUs," IEEE Transactions on Big Data, vol. 7, no. 3, pp. 535–547, 2019.

[9] C. Jia, Y. Yang, Y. Xia, et al., "Scaling up visual and vision-language representation learning with noisy text supervision," in International Conference on Machine Learning (ICML), 2021.

[10] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese BERT-networks," in Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2019.

[11] Z. Lin, M. Feng, C. N. dos Santos, et al., "A structured self-attentive sentence embedding," in International Conference on Learning Representations (ICLR), 2017.

[12] R. Caruana, "Multitask learning," Machine Learning, vol. 28, no. 1, pp. 41–75, 1997.

[13] A. Kendall, Y. Gal, and R. Cipolla, "Multi-task learning using uncertainty to weigh losses for scene geometry and semantics," in Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2018.

[14] Z. Chen, V. Badrinarayanan, C.-Y. Lee, and A. Rabinovich, "Gradnorm: Gradient normalization for adaptive loss balancing in deep multitask networks," in International Conference on Machine Learning (ICML), 2018.

[15] A. Conneau, K. Khandelwal, N. Goyal, et al., "Unsupervised cross-lingual representation learning at scale," in Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), 2020.

[16] J. L. Ba, J. R. Kiros, and G. E. Hinton, "Layer normalization," arXiv preprint arXiv:1607.06450, 2016.

[17] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," in International Conference on Learning Representations (ICLR), 2019.

[18] T.-Y. Lin, M. Maire, S. Belongie, et al., "Microsoft COCO: Common objects in context," in European Conference on Computer Vision (ECCV), 2014.

[19] P. Young, A. Lai, M. Hodosh, and J. Hockenmaier, "From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions," Transactions of the Association for Computational Linguistics, vol. 2, pp. 67–78, 2014.

[20] T. A. Nguyen et al., "A comprehensive benchmark for Vietnamese text evaluation," in Proc. VLSP, 2023.

[21] M. Mathew, R. Karatzas, and C. V. Jawahar, "DocVQA: A dataset for VQA on document images," in Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV), 2021.

[22] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 2020.

[23] W.-N. Hsu, B. Bolte, Y.-H. H. Tsai, et al., "HuBERT: Self-supervised speech representation learning by masked prediction of hidden units," IEEE/ACM Transactions on Audio, Speech, and Language Processing, vol. 29, pp. 3451-3460, 2021.

[24] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "AudioCLIP: Extending CLIP to image, text and audio," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.

[25] H. Wang, Y. Zhang, Z. Li, et al., "CLAP: Learning audio concepts from natural language supervision," in IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2023.

[26] K. Drossos, S. Lipping, and T. Virtanen, "Clotho: An audio captioning dataset," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2020.

[27] C. D. Kim, B. Kim, H. Lee, and G. Kim, "AudioCaps: Generating captions for audios in the wild," in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019.

[28] A. Guzhov, F. Raue, J. Hees, and A. Dengel, "AudioCLIP: Extending CLIP to image, text and audio," in IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), 2022.