Okay, here is a draft for a 2-page ArXiv-style paper based on the provided code and information. It emphasizes the technical details, justifications, and potential advantages, while acknowledging the current lack of benchmarks.

---

viPolyQwen: Unified Multimodal Embeddings via Prefix-Guided Dynamic Loss and Attention Pooling

Steve Nguyen Anh Nguyen*, **EraX AI Team**, **AI Technology Team, Gtel Mobile JSC (GMobile)** * Corresponding Author: nguyen@hatto.com

(Draft - Work in Progress)

Abstract

Effectively representing diverse multimodal data (text, images, documents) within a unified vector space is crucial for applications like Retrieval-Augmented Generation (RAG) and cross-modal search, yet challenging. Existing methods often rely on separate embeddings or simple pooling strategies, potentially limiting cross-modal understanding and nuance. We introduce ***viPolyQwen****, a multimodal embedding model designed to generate a single, high-dimensional (1024-d) vector representation for varied inputs. Built upon the Qwen2-VL-2B-Instruct architecture, viPolyQwen leverages a large-scale (>11M samples), diverse dataset encompassing text similarity, instructions, OCR, and multi-turn VQA tasks, with a strong focus on Vietnamese alongside English and Chinese data. Training employs a novel* ***prefix-guided dynamic mixed-loss optimization*** *strategy, where task-specific prefixes trigger tailored contrastive loss functions (InfoNCE, Triplet, MSE, Cosine). Crucially, the final 1D embedding is derived via* ***Attention Pooling****, allowing the model to dynamically weight salient features from the encoder's output sequence, generating richer representations compared to mean or last-token pooling, especially for inputs like text-rich images. This unified, attention-pooled embedding potentially offers a simpler yet powerful alternative to multi-vector approaches for complex multimodal retrieval and analysis.*

1. Introduction

The proliferation of multimodal data necessitates models capable of understanding and relating information across modalities like text and images. Dense retrieval systems, particularly for RAG, require high-quality embeddings that capture semantic and visual essence within a computationally tractable format. While Vision-Language Models (VLMs) like CLIP [Radford et al., 2021] and its successors have advanced cross-modal understanding, generating effective *task-agnostic* yet *task-aware* embeddings for diverse downstream applications remains an open challenge. Approaches often involve separate embedding spaces or multi-vector representations [Faysse et al., 2024], which can increase system complexity for indexing and retrieval. Furthermore, standard pooling techniques like mean or last-token pooling applied to VLM encoder outputs might average out or ignore critical features, especially in information-dense inputs like documents or images containing text.

To address these limitations, we propose **viPolyQwen**, a model aiming to produce a single, unified 1D embedding vector (1024-d) for diverse multimodal inputs. Our primary contributions are:

1. **Unified 1D Embedding Space:** Generating a single vector for text, images, and combinations, simplifying downstream integration.
2. **Prefix-Guided Dynamic Loss:** A training paradigm using task prefixes (`<text_pair>`, `<instr>`, `<ocr>`, `<vqa_...>`) to dynamically select optimal contrastive loss functions (InfoNCE, Triplet, MSE, Cosine Similarity) based on the input data type during training.
3. **Attention Pooling Mechanism:** Employing a learnable attention mechanism over the VLM encoder's final hidden states sequence to compute a weighted average, focusing on salient features and producing more nuanced 1D embeddings than traditional pooling.
4. **Diverse Training Data & Vietnamese Focus:** Training on a large (>11M), heterogeneous dataset including similarity, instructions, complex OCR/VQA (documents, medical images), with emphasis on Vietnamese alongside multilingual data for zero-shot potential.
5. **Potential Simplification:** Offering a potentially simpler alternative to multi-vector approaches [Faysse et al., 2024] for building powerful multimodal retrieval systems.

This work was developed in collaboration with the AI technology team at Gtel Mobile JSC (GMobile).

2. Related Work

Multimodal representation learning has seen significant progress, largely driven by contrastive learning on image-text pairs (e.g., CLIP [Radford et al., 2021], ALIGN [Jia et al., 2021]). Fine-tuning VLMs for specific embedding tasks, like text embedding (e.g., Sentence-BERT [Reimers & Gurevych, 2019] adapted for multimodal contexts) or retrieval, is common. However, creating a single embedding space that handles diverse task structures (similarity, instruction following, OCR, VQA) effectively remains challenging.

Recently, models like ColPali [Faysse et al., 2024] proposed multi-vector representations for documents, using separate vectors for global context and local patches, requiring specialized retrieval mechanisms (e.g., ColBERT-style Late Interaction). While potentially capturing fine-grained details, this adds complexity. Our work explores the alternative hypothesis: can a sufficiently powerful VLM, trained with dynamic task-aware losses and a sophisticated pooling mechanism like Attention Pooling, generate a *single* 1D vector rich enough for diverse multimodal tasks, thereby simplifying system design?

3. Methodology

3.1 Model Architecture

viPolyQwen builds upon the `Qwen/Qwen2-VL-2B-Instruct` [Bai et al., 2023] VLM. The core embedding generation process (detailed in `model (13).py`) follows these steps:

1. **Input Processing:** Text and images are processed using the Qwen-VL processor. Task prefixes are prepended to text inputs *during training only* as per the data type.
2. **Multimodal Encoding:** The Qwen-VL encoder processes the tokenized text and image patches, outputting a sequence of final hidden states $\mathbf{H} = [\mathbf{h}_1, \mathbf{h}_2, ..., \mathbf{h}_N] \in \mathbb{R}^{N \times D_{hidden}}$, where N is the sequence length and D_{hidden} is the hidden dimension of the base VLM. These hidden states represent both text tokens and processed visual features.
3. **Attention Pooling:** Instead of mean or last-token pooling, we apply Attention Pooling (Sec 4.1) to $\mathbf{H}$ to obtain a single context vector $\mathbf{c} \in \mathbb{R}^{D_{hidden}}$.
4. **Projection & Normalization:** The pooled vector $\mathbf{c}$ is passed through a projection head (`self.proj`), consisting of a linear layer followed by Layer Normalization: $\mathbf{p} = \text{LayerNorm}(\mathbf{W}_{proj}\mathbf{c})$, where $\mathbf{W}_{proj} \in \mathbb{R}^{D_{embed} \times D_{hidden}}$ and $D_{embed} = 1024$.
5. **Final Embedding:** The projected vector $\mathbf{p}$ is L2-normalized to produce the final embedding: $\mathbf{e} = \mathbf{p}/||\mathbf{p}||_2$.

3.2 Training Paradigm

Training leverages a large (>11M samples), diverse dataset sourced from various public and private collections, covering: * Text similarity pairs (Vietnamese, English, Chinese) with scores. * Instruction-following data (text-only and multimodal). * OCR/OCQ data from documents, receipts, handwriting. * Single and multi-turn VQA data, including general knowledge, document/chart analysis, and specialized medical image QA.

The key innovation is the **Prefix-Guided Dynamic Mixed-Loss Optimization** (implemented in `train (9).py`, `mix_data_collator (5).py`, `losses (6).py`). Each training sample is prepended with a task-specific prefix (e.g., `<ocr>`). The `multi_purpose_contrastive_loss` function (Sec 4.2) uses this prefix to dispatch the calculation to a tailored loss function operating on the final L2-normalized embeddings ($\mathbf{e}_a, \mathbf{e}_b$) derived via Attention Pooling. Training was performed on 4x H100 GPUs using FSDP and `bfloat16` precision (see `README_en.md` for full hyperparameters).

4. Key Mechanisms

4.1 Attention Pooling

Given the sequence of final hidden states $\mathbf{H} = [\mathbf{h}_1, ..., \mathbf{h}_N]$, Attention Pooling computes the summary vector $\mathbf{c}$ as follows:

1. **Learnable Context:** A learnable parameter vector $\mathbf{v}_a \in \mathbb{R}^{D_{hidden}}$ is introduced, representing a task-agnostic "query" for importance.
2. **Attention Scores:** Unnormalized attention scores e_i are computed for each hidden state $\mathbf{h}_i$: $e_i = \mathbf{h}_i^T \mathbf{v}_a$

3. **Masking:** Scores corresponding to padding tokens (identified by the attention mask) are set to $-\infty$.
4. **Attention Weights:** Scores are normalized using softmax: $\alpha_i = \frac{\exp(e_i)}{\sum_{j=1}^{N}\exp(e_j)}$
5. **Weighted Average:** The final pooled vector $\mathbf{c}$ is the weighted sum: $\mathbf{c} = \sum_{i=1}^{N}\alpha_i \mathbf{h}_i$

Justification: Unlike mean pooling (uniform weighting) or last-token pooling (ignores prior context), Attention Pooling learns to dynamically assign higher weights (α_i) to hidden states ($\mathbf{h}_i$) deemed more relevant, guided by the learned context $\mathbf{v}_a$. This allows the model to focus on salient features (e.g., keywords, specific visual regions, text within images) when creating the summary vector $\mathbf{c}$, leading to potentially richer and more nuanced 1D embeddings crucial for capturing the essence of complex inputs.

4.2 Dynamic Loss Function

The core training loss is computed via `multi_purpose_contrastive_loss` (defined in `losses (6).py`). For a batch of embedding pairs $(\mathbf{e}_{a,i}, \mathbf{e}_{b,i})$ and corresponding data types $type_i$, the total loss is an average over specialized loss functions $\mathcal{L}_{type}$:

$\mathcal{L}_{total} = \frac{1}{B}\sum_{i=1}^{B}\mathcal{L}_{type_i}(\mathbf{e}_{a,i}, \mathbf{e}_{b,i}, \text{params})$

Where `params` include temperature T, margin m, and potential similarity scores s_i. Key loss components include:

- **Symmetric InfoNCE:** $\mathcal{L}_{NCE}(\mathbf{e}_a, \mathbf{e}_b) = -\frac{1}{2B}\sum_{i=1}^{B}[\log \frac{\exp(sim(\mathbf{e}_{a,i},\mathbf{e}_{b,i})/T)}{\sum_{j=1}^{B}\exp(sim(\mathbf{e}_{a,i},\mathbf{e}_{b,j})/T)} + \log \frac{\exp(sim(\mathbf{e}_{b,i},\mathbf{e}_{a,i})/T)}{\sum_{j=1}^{B}\exp(sim(\mathbf{e}_{b,i},\mathbf{e}_{a,j})/T)}]$
- **MSE Similarity Regression:** (for `<text_pair>`) $\mathcal{L}_{MSE} = \frac{1}{B}\sum_{i=1}^{B}(\frac{sim(\mathbf{e}_{a,i},\mathbf{e}_{b,i})+1}{2} - s_i)^2$
- **Direct Cosine Similarity:** (for `<instr>`) $\mathcal{L}_{Cos} = 1 - \frac{1}{B}\sum_{i=1}^{B} sim(\mathbf{e}_{a,i}, \mathbf{e}_{b,i})$
- **Triplet Margin Loss:** (for `<ocr>`, `<vqa_...>`, uses scaled similarities) $\mathcal{L}_{Triplet} = \frac{1}{B}\sum_{i=1}^{B}\max(0, \max_{j\neq i}\frac{sim(\mathbf{e}_{a,i},\mathbf{e}_{b,j})}{T} - \frac{sim(\mathbf{e}_{a,i},\mathbf{e}_{b,i})}{T} + m)$

The specific combination (e.g., $\mathcal{L}_{NCE} + \mathcal{L}_{MSE}$ for text pairs) is chosen based on the `data_type` prefix.

5. Potential Advantages & Discussion

While comprehensive benchmarks are pending, the design of viPolyQwen offers potential advantages, particularly compared to multi-vector approaches like ColPali [Faysse et al., 2024]:

- **System Simplicity:** A single 1024-d vector per item significantly simplifies indexing (standard vector DBs suffice) and retrieval (single similarity computation vs. multi-stage scoring). This can reduce engineering overhead and potentially inference latency.

- **Unified Representation Power:** We hypothesize that the combination of a powerful base VLM, diverse task-aware training, and sophisticated Attention Pooling allows the model to encode rich multimodal information, including text-in-image details and cross-modal relationships, within a single vector. Attention Pooling is key here, as it avoids information dilution inherent in mean pooling.
- **Implicit Cross-Modal Interaction:** Training diverse data types towards a unified embedding space might encourage the model to learn stronger implicit correlations between modalities compared to systems managing separate representations.
- **Versatility:** The single embedding is directly usable for various tasks: semantic search, visual search, cross-modal retrieval, clustering, and potentially as input features for downstream classifiers, without requiring task-specific heads during inference (except when using task-specific prefixes for querying, see `USAGE.md`).

However, this approach relies heavily on the capacity of the 1024-d vector and the effectiveness of Attention Pooling to capture sufficient detail. Multi-vector approaches may still hold advantages in scenarios requiring extremely fine-grained localization or interaction. Empirical validation is crucial to determine the trade-offs.

6. Conclusion & Future Work

viPolyQwen presents a promising approach towards unified multimodal embeddings, leveraging prefix-guided dynamic loss optimization and Attention Pooling on a strong VLM foundation. By generating a single, nuanced 1D vector, it aims to simplify the architecture of multimodal retrieval systems while maintaining high representational power, particularly for Vietnamese and potentially zero-shot cross-lingual tasks.

Future work will focus on: 1. **Comprehensive Benchmarking:** Evaluating viPolyQwen on standard Vietnamese and English multimodal retrieval, classification, and STS tasks, comparing against relevant baselines including multi-vector methods. 2. **Ablation Studies:** Quantifying the impact of Attention Pooling vs. other pooling methods and the contribution of different data types/loss components. 3. **Exploration of Base Models:** Adapting the framework to larger or different VLM architectures.

The model and evaluation code will be released upon completion of benchmarking.

References

[Bai et al., 2023] Jinze Bai, Shuai Bai, et al. *Qwen-VL: A Versatile Vision-Language Model for Understanding, Localization, Text Reading, and Beyond.* arXiv:2308.12966, 2023.

[Faysse et al., 2024] Manuel Faysse, Hugues Sibille, et al. *ColPali: Efficient Document Retrieval with Vision Language Models.* arXiv:2407.01449, 2024.

[Jia et al., 2021] Chao Jia, Yinfei Yang, et al. *Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision.* ICML 2021.

[Radford et al., 2021] Alec Radford, Jong Wook Kim, et al. *Learning Transferable Visual Models From Natural Language Supervision.* ICML 2021.

[Reimers & Gurevych, 2019] Nils Reimers and Iryna Gurevych. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks.* EMNLP 2019.

[ViPolyQwen Repo, 2024] Steve Nguyen Anh Nguyen, et al. *viPolyQwen GitHub Repository.* https://github.com/EraX-AI/viPolyQwen, 2024.

(Additional references for PyTorch, Transformers, etc., would be added)