FrEVL: Leveraging Frozen Pretrained Embeddings for Efficient Vision-Language Understanding Supplementary Material

A. Theoretical Analysis of Representation Sufficiency

To understand when frozen embeddings suffice for downstream tasks, we analyze the information-theoretic properties of our approach. Let $\mathcal{H}(Y|V,T)$ denote the conditional entropy of task labels given perfect visual and textual information, and $\mathcal{H}(Y|\mathbf{v},\mathbf{t})$ denote the conditional entropy given only frozen embeddings.

For a downstream task with label space $\mathcal{Y}$, the performance gap between using full representations and frozen embeddings is bounded by:

$$\Delta_{\text{perf}} \leq C \cdot [\mathcal{H}(Y|\mathbf{v},\mathbf{t}) - \mathcal{H}(Y|V,T)] \tag{1}$$

where C depends on the task loss function and model capacity.

Let $f^* : \mathcal{V} \times \mathcal{T} \to \mathcal{Y}$ be the optimal predictor using full representations, and $g^* : \mathbb{R}^{d_v} \times \mathbb{R}^{d_t} \to \mathcal{Y}$ be the optimal predictor using frozen embeddings. The performance gap can be expressed as:

$$\Delta_{\text{perf}} = \mathbb{E}[\ell(g^*(\mathbf{v},\mathbf{t}),Y)] - \mathbb{E}[\ell(f^*(V,T),Y)] \tag{2}$$

$$\leq \mathbb{E}[|g^*(\mathbf{v},\mathbf{t}) - f^*(V,T)|] \tag{3}$$

$$\leq \sqrt{\text{Var}[Y|\mathbf{v},\mathbf{t}] - \text{Var}[Y|V,T]} \tag{4}$$

Using the relationship between variance and entropy for discrete distributions with bounded support:

$$\text{Var}[Y|X] \leq K \cdot \mathcal{H}(Y|X) \tag{5}$$

where K depends on the label space cardinality. Combining these inequalities yields the desired bound.

This bound reveals that performance degradation depends on how much task-relevant information is lost during the encoding process. For tasks where embeddings preserve most discriminative information (e.g., semantic similarity), the gap is small. However, for tasks requiring information not captured during pretraining (e.g., counting, OCR), the gap can be arbitrarily large. This theoretical insight explains our empirical findings and provides guidance on task suitability.

Implications for Pretraining Objectives. The theorem suggests that improving frozen embedding approaches requires pretraining objectives that minimize $\mathcal{H}(Y|\mathbf{v},\mathbf{t})$ for a wide range of downstream tasks Y. Current contrastive objectives optimize for image-text alignment but may discard information crucial for other tasks. Future work might explore multi-task pretraining or information-maximizing objectives that preserve more diverse task-relevant signals in the final embeddings.

B. Implementation Details

Hardware and Software Configuration. All experiments were conducted on NVIDIA V100 16GB GPUs. We used PyTorch 2.0.1 with CUDA 11.8, transformers 4.35.0 for baseline models, OpenCLIP 2.20.0 for our encoder variants. For efficiency purposes, we used mixed precision training using PyTorch AMP and gradient checkpointing.

Training Hyperparameters. Table 1 shows the detailed hyperparameters for each dataset. The learning rates range from 2e-5 to 1e-4 depending on the dataset complexity, with warmup steps varying from 500 to 2000. Training epochs were set between 20 and 40 epochs based on convergence behavior, and batch sizes were either 256 or 512 depending on memory constraints and dataset size.

Table 1. Dataset-specific training configurations

Dataset	LR	Warmup	Epochs	Batch Size
COCO	1e-4	1000	30	512
VQA v2	5e-5	2000	20	512
SNLI-VE	1e-4	1500	30	512
MMMU	2e-5	500	40	256
MMBench	2e-5	500	40	256

Architecture Details. Our fusion network architecture consists of several key components. The projection layers perform linear transformation from the embedding dimension of 768 for CLIP-L to a hidden dimension of 512,

followed by GELU activation and LayerNorm. The cross-attention blocks contain 4 transformer layers with 8 attention heads at 64 dimensions per head, FFN hidden dimension of 2048 representing a 4× expansion, pre-LayerNorm configuration, and dropout of 0.1 on both attention and FFN. The fusion layer concatenates $[\mathbf{v}, \mathbf{t}, \mathbf{v} \odot \mathbf{t}, |\mathbf{v} - \mathbf{t}|]$ to produce 2048-dimensional features. Finally, the prediction head consists of a two-layer MLP that transforms from 2048 to 1024 dimensions and then to the output dimension, using GELU activation and dropout of 0.1.

Data Preprocessing. Images are preprocessed using CLIP's standard pipeline. We first resize images to 224×224 using bicubic interpolation, normalize with CLIP statistics. We use random horizontal flip during training (except for VQA v2 dataset). We do not employ additional augmentation to preserve embedding quality. Text preprocessing follows CLIP tokenization with maximum sequence length of 77 tokens, and lowercase normalization.

Embedding Storage and Caching. For efficiency, we implement a caching system. We compute the pre-computed embeddings which are stored in HDF5 format with compression, reducing storage from 6KB to 2KB per sample while maintaining numerical precision.

C. Human Evaluation Details

Table 2. Human evaluation with inter-rater agreement

Method	Quality (1-5)		Preference		Agreement	
	Relevance	Coherence	Win%	Lose%	κ	α
Human	4.68±0.52	4.71±0.48	-	-	0.856	0.871
BLIP	4.31±0.61†	4.38±0.57†	48.3	39.2	0.823	0.841
FrEVL-B	4.19±0.65	4.28±0.61	42.9	44.6	0.834	0.848
Random	2.43±0.89†	2.16±0.92†	8.7	78.4	-	-

†Significant vs FrEVL ($p < 0.05$). κ: Cohen's kappa, α: Krippendorff's alpha

Inter-Annotator Agreement. Table 2 demonstrates high inter-annotator agreement for FrEVL outputs with Cohen's $\kappa = 0.834$ and Krippendorff's $\alpha = 0.848$, indicating consistent quality. The agreement levels match those for full model outputs, suggesting that frozen embedding approaches do not produce more ambiguous or inconsistent results despite their architectural constraints.

Agreement analysis by task type reveals varying levels of consensus across different datasets. The highest agreement was observed on SNLI-VE with $\kappa = 0.867$ due to clear entailment decisions, moderate agreement on VQA with $\kappa = 0.821$ reflecting some answer ambiguity, and lower agreement on COCO with $\kappa = 0.798$ due to subjective caption quality assessments.

Qualitative Patterns. Manual analysis of outputs reveals clear patterns in success and failure cases. FrEVL excels when tasks require semantic matching, scene understanding, or general object recognition. Failures concentrate on counting ("three dogs" → "dogs"), spatial relationships ("cat on the left of the dog" → "cat and dog"), text in images (missing store signs), and fine-grained attributes ("spotted dalmatian" → "dog"). These patterns align perfectly with our theoretical analysis of what information frozen embeddings can and cannot capture.

Error Analysis. Categorizing 500 error cases from human evaluation reveals distinct failure modes. Missing details account for 38% of errors, involving omitted specific attributes, counts, or fine-grained information. Spatial errors comprise 24% of failures with incorrect or missing spatial relationships. Hallucination represents 18% of errors where the model adds information not present in the image. Misalignment accounts for 12% of cases with correct information but poor relevance to the query. The remaining 8% consists of other errors including grammatical mistakes and incomplete responses.

Annotation Guidelines. Annotators were provided with detailed guidelines covering three key criteria. For relevance, annotators assessed how well the output addresses the input query or task. For coherence, they evaluated whether the output is grammatically correct and logically consistent. For preference, they determined which output they would prefer in a real application. Training included 100 calibration examples with discussion to ensure consistent standards.

D. Detailed Efficiency Analysis

Memory-Constrained Environments. On edge devices with 4GB memory limits, FrEVL's 2.3GB footprint enables deployment where 8.7GB+ full models cannot run at all. For batch processing scenarios, FrEVL maintains efficient memory scaling up to batch size 256 on 16GB GPUs, compared to batch size 64 limits for full models. This 4× throughput advantage compounds in production systems handling high request volumes.

The memory breakdown for FrEVL-B consists of 274MB for model parameters calculated as 68.4M × 4 bytes, 896MB for activation memory with batch size 32, approximately 1.1GB for PyTorch overhead, resulting in a total memory footprint of 2.3GB.

Energy and Cost Analysis. Table 3 presents detailed energy measurements over a 24-hour period. In 24/7 deployment scenarios, energy savings translate directly to operational costs. At 0.12 per kWh, FrEVL saves approximately 11,400 annually per deployment compared to full models. For organizations running hundreds of instances, this represents millions in reduced operational expenses. The environmental impact is equally significant, with each deployment saving 73.6 tons of CO_2 annually using US grid emission factors.

Table 3. Energy consumption over 24-hour period

Method	Avg Power	Total kWh	Annual Cost
BLIP	389W	9.34	$410
FrEVL-B	187W	4.49	$197

Latency-Sensitive Applications. For real-time applications, FrEVL's consistent 2ms inference latency after embedding extraction enables deployment in interactive systems. The predictable performance characteristics, without the variable latency of autoregressive generation, simplify system design and capacity planning. However, applications requiring dynamic visual inputs must account for embedding extraction time, reducing effective speedup to 2.3× rather than theoretical maximums.

The latency breakdown shows 18ms for embedding extraction using CLIP-L/14, 2ms for fusion network forward pass, resulting in 20ms total end-to-end latency compared to 46ms for the full BLIP model.

Pre-Computation Opportunities. Many real-world scenarios allow embedding pre-computation across various domains. E-commerce platforms with fixed product catalogs can pre-compute embeddings during catalog updates. For a 1M product catalog, storage of 6GB can be compressed to 2GB with update time of 4.2 hours on a single GPU and query latency of just 2ms for fusion only. Content moderation systems can embed user-uploaded media during upload, adding only 18ms processing overhead to uploadss while achieving 2ms moderation latency and 500 decisions per second throughput. Educational applications with curated content libraries benefit from one-time setup for course materials, enabling instant response for student queries and scaling to thousands of concurrent users.

Scaling Analysis. Table 4 shows performance characteristics across deployment scales. The consistent 2.5-3× speedup across hardware platforms demonstrates FrEVL's architectural efficiency rather than optimization for specific accelerators.

Table 4. Scaling comparison with hardware

Hardware	FrEVL FPS	BLIP FPS	Speedup
V100 (16GB)	412	156	2.6×
A100 (40GB)	512	172	3.0×
8×A100 cluster	3,847	1,243	3.1×
TPU v4	892	298	3.0×