
11-785 Spring 2025 Project Midterm Report - Group 35

Yuanyu Wang, Zhihan Li
{helenwan, zhihan13}@andrew.cmu.edu

Abstract

Recent advancements in vision-language modeling have relied heavily on token-level representations [Radford et al., 2021, Alayrac et al., 2022, Li et al., 2022], which limit generalization across languages and constrain models’ ability to perform high-level reasoning. Moreover, token-based approaches introduce a fundamental mismatch: while text consists of discrete tokens, images do not naturally decompose into equivalent units. In many tasks, a single image semantically corresponds to an entire sentence—not a sequence of tokens—suggesting that sentence-level representations offer a more aligned abstraction.

In this work, we propose aligning visual features with the SONAR sentence-level embedding space [Duquenne et al., 2023], a modality-agnostic space designed for multilingual, conceptual reasoning. We fine-tune a CLIP vision encoder using an image captioning task on the COCO dataset [Lin et al., 2015], minimizing cosine similarity between mapped image embeddings and SONAR text embeddings. A lightweight MLP projects the 512-dimensional CLIP outputs to the 1024-dimensional SONAR space, enabling caption generation via SONAR’s pretrained decoder. Our framework avoids token-level supervision and supports zero-shot captioning. Despite its simplicity, the model achieves performance competitive with decoder-based baselines like BLIP and BLIP-2 [Li et al., 2023], supporting the potential of concept-based, language-agnostic embeddings in vision-language systems.

1 Introduction

Motivation Existing vision-language models such as CLIP Radford et al. [2021] and Flamingo Alayrac et al. [2022] align vision and language by mapping them into a shared embedding space, but they predominantly rely on token-level processing. This approach introduces limitations, particularly for tasks that require deeper reasoning or cross-lingual generalization. These models are often trained and evaluated on English-only benchmarks, which makes their embeddings highly language-dependent and restricts transfer to low-resource languages.

Moreover, token-level processing restricts the ability of these models to capture high-level semantic concepts. Human cognition naturally operates at a more abstract level, processing ideas and meanings rather than isolated words. This limitation means that current models often fail to fully encapsulate the deeper, language-agnostic understanding necessary for effective cross-modal reasoning and alignment.

Motivated by these challenges, our research proposes integrating Large Concept Models (LCMs) Barrault et al. [2024] into the vision-language framework by leveraging the SONAR embedding space, which produces sentence-level, multilingual embeddings designed for abstraction and generalization. By moving from token-based representations to a concept-based approach, we aim to achieve a modality-agnostic framework that improves abstraction, reasoning, and cross-modal alignment. This

shift not only mitigates language dependency but also supports robust zero-shot performance across languages and modalities.

Objective The goal of this work is to evaluate the effectiveness of SONAR’s embedding space Duquenne et al. [2023] in vision-language modeling tasks. Specifically, we want to determine whether a pretrained vision encoder can be adapted to produce embeddings that align well with SONAR’s concept-based space.

To explore this, we fine-tune the vision encoder using an image captioning task on the COCO dataset Lin et al. [2015]. During training, we minimize cosine similarity loss between the embeddings generated by the vision encoder and those produced by the SONAR text encoder. Since SONAR embeddings are 1024-dimensional and the visual encoder outputs 512-dimensional features, we introduce a vision mapper, implemented as a simple MLP, to transform the visual features into the SONAR space.

Our objectives are:

- **Qualitative:** To explore whether concept-level alignment enables coherent caption generation and captures high-level semantics from images.
- **Quantitative:** To evaluate the alignment quality using cosine similarity loss and standard captioning metrics such as BLEU-4 and CIDEr, providing evidence of the model’s generalization.

2 Related Work

Sentence-level Embedding Space The Large Concept Model (LCM) Barrault et al. [2024] shifts away from token-based models by operating on fixed-size semantic representations—“concepts”—that capture the meaning of entire sentences. This abstraction mirrors human reasoning and supports tasks like summarization and translation across multiple languages and modalities. At the core of LCM is the SONAR embedding space Duquenne et al. [2023], which encodes text and speech segments into robust, fixed-size vectors across over 200 languages and 76 speech languages. SONAR’s encoder-decoder strategy, using objectives like translation and auto-encoding, yields semantically rich embeddings effective for cross-lingual similarity and machine translation. Combining LCM with SONAR decouples high-level reasoning from language-specific details, enabling document segmentation into concept embeddings that can be decoded into text or other modalities, thus enhancing long-form coherence and zero-shot generalization.

Vision Language Models Vision-language tasks require models to integrate visual and textual information to perform challenges like image captioning, visual question answering (VQA), and cross-modal retrieval. These tasks demand capturing fine-grained visual details and understanding linguistic nuances, which calls for robust joint representations. Recent advances in this field have been driven by large-scale pre-training and transformer-based architectures. Notably, CLIP Radford et al. [2021] and BLIP Li et al. [2022] have emerged as influential models. CLIP uses contrastive learning on millions of image-text pairs to align visual and textual representations in a shared space, enabling impressive zero-shot classification and retrieval. In contrast, BLIP introduces a bootstrapping mechanism that improves image-text alignment using caption refinement and question-answer generation.

Building on these developments, BLIP-2 Li et al. [2023] introduces a two-stage approach that first aligns vision and language through a lightweight Q-former and then performs reasoning via a frozen language model. This modular design improves sample efficiency and allows for leveraging powerful pretrained language models without full-scale joint training. Another notable contribution is Flamingo Alayrac et al. [2022], which introduces a few-shot vision-language model that fuses frozen vision encoders with pretrained language models via gated cross-attention layers. Flamingo achieves strong performance on a wide range of multimodal benchmarks with minimal task-specific finetuning. These models highlight the trend toward modular and efficient architectures that decouple vision and language processing while maintaining strong alignment between modalities.

3 Methodology

Model Description We introduce an SONAR-embedding-based Vision-Language Model (SONAR-VLM) with the following architectures.

3.1 System Overview

Our model consists of four main components: a **vision encoder**, the **SONAR text encoder** Duquenne et al. [2023], the **SONAR text decoder** Duquenne et al. [2023], and a **vision-to-text mapping layer**. The vision encoder extracts high-level features from input images, which are transformed by the mapping layer into the SONAR embedding space. These embeddings are then passed to the SONAR text decoder to generate descriptive captions.

3.2 Model Description

The architecture integrates a vision transformer (ViT-B/16 or ViT-B/32) as the visual backbone. The models produce 512-dimensional embeddings for images, which are projected into SONAR’s 1024-dimensional language embedding space using a multi-layer perceptron (MLP) projection layer. The projection layer consists of two linear layers with a ReLU activation in between.

Table 1: Model architecture and dimensions

Component	Specification
Vision Encoder	CLIP ViT-B/16 or ViT-B/32, 512-dim output
Vision Mapper	MLP: Linear(512 $\rightarrow$ 1024), ReLU, Linear(1024 $\rightarrow$ 1024)
Text Encoder	SONAR encoder, 1024-dim output
Text Decoder	SONAR decoder

We employ the encoder and decoder modules from the SONAR language model. During training, we adopt a “teacher-student” approach Reimers and Gurevych [2020], Heffernan et al. [2022] for training the image encoder. The image embeddings are aligned with the SONAR language embeddings, and caption generation is performed by decoding the aligned representations of the image encoder. In this framework, the pre-trained text encoder functions as the “teacher,” providing sentence-level semantic representations, while the image encoder is the “student” that learns to map visual features into the same embedding space. This method leverages the robust, language-agnostic properties of SONAR, ensuring that the image-derived embeddings are semantically aligned with textual representations, which is crucial for cross-modal tasks.

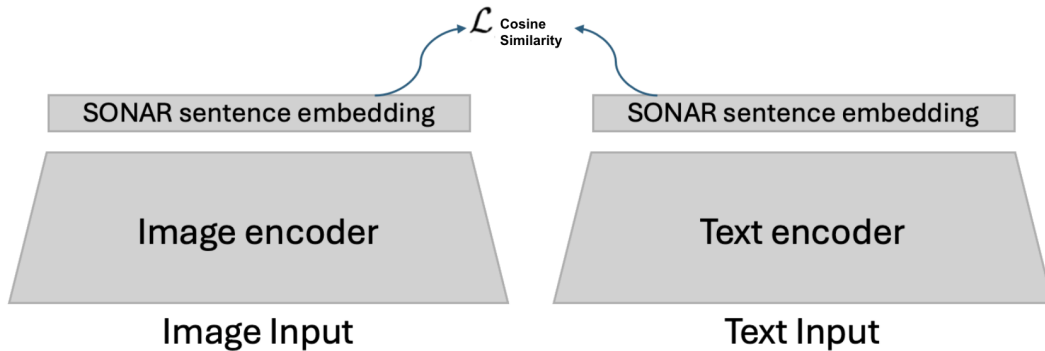


Figure 1: Training the image-concept encoder

Training Objective Although our broader aim is to build a general-purpose vision-language model (VLM) that maps images directly into a semantically meaningful embedding space (similar to CLIP), we train the model using a supervised image captioning task on the MS-COCO dataset. Specifically, we minimize cosine similarity loss between the mapped visual embedding and the SONAR text

encoder output for the corresponding caption. This enables the model to learn alignment between image content and language in a shared vector space.

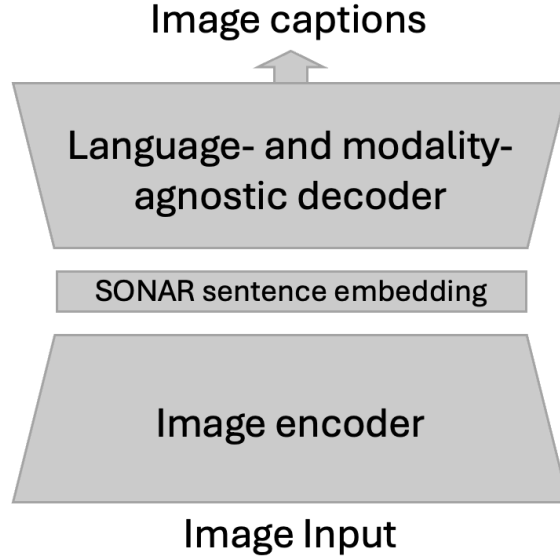


Figure 2: Setup to generate image captions

This setup is conceptually similar to how many language models are trained on proxy objectives like next-token prediction, yet are ultimately used for tasks such as translation or summarization. In our case, while training is done via caption alignment, the learned image-to-embedding mapping is intended for use in downstream tasks that benefit from a shared vision-language space (e.g., retrieval, classification, or captioning via decoding).

Token-Independent Concept Inference: Our approach enables the inference of conceptual meaning directly from the SONAR sentence embedding space without relying on predefined textual tokens. Once the image-to-sentence encoder is trained, image captioning becomes straightforward: the produced SONAR embedding can be directly decoded into a readable sentence using a language- and modality-agnostic decoder. Moreover, by leveraging a speech decoder, the same embedding can be transformed into speech, demonstrating the modality flexibility of our framework.

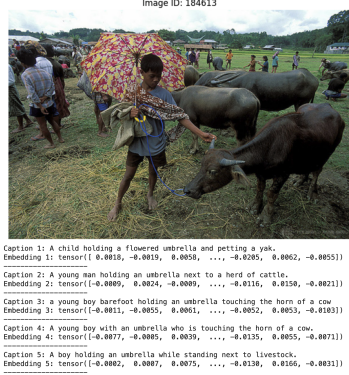
Furthermore, this token-independent approach extends to tasks like Visual Question Answering (VQA). By concatenating the SONAR embeddings derived from both an image and its associated question, we can train a dedicated VQA model that operates in the high-level concept space. This strategy bypasses the limitations of token-level processing, enabling robust cross-modal reasoning and improved generalization across languages and modalities.

3.3 Dataset

We use the COCO 2017 dataset Lin et al. [2015] for training and evaluation, with `train2017` (around 118K images) for training and `val2017` (around 5K images) for validation. Image preprocessing follows the CLIP pipeline. Captions are embedded using SONAR’s language encoder. The model is trained using mini-batches of size 64, and sampling is randomized each epoch to ensure diverse coverage.

3.4 Evaluation Metric

To evaluate the model’s alignment and captioning performance, we use two types of metrics: **BLEU-4**. We use BLEU-4 (Bilingual Evaluation Understudy with up to 4-grams) to measure n-gram overlap between the generated caption and one or more human-written references. The BLEU score is computed as the geometric mean of the modified n-gram precisions, multiplied by a brevity penalty (BP) to penalize overly short outputs:



(a) Captions and SONAR embeddings for image 184613

BLIP Caption: a clock mounted on the back of a bike
GT Caption: A clock with the appearance of the wheel of a bicycle



(b) BLIP vs. Ground Truth Caption on COCO2017

Figure 3: Comparison of SONAR embeddings (left) and BLIP-generated vs. Ground Truth captions (right).

$$BLEU - N = BP \cdot \exp \left(\sum_{n=1}^N w_n \cdot \log p_n \right) \quad (1)$$

$$BP = \{ 1 \text{ if } c > r \exp(1 - \frac{r}{c}) \text{ if } c \leq r \quad (2)$$

where p_n is the modified precision for n-grams, w_n is typically $\frac{1}{N}$, and c, r represent candidate and reference lengths.

CIDEr

CIDEr (Consensus-based Image Description Evaluation) computes the similarity between a candidate caption and its reference set using TF-IDF weighted n-gram vectors. For each n-gram size, the CIDEr score is computed as:

$$CIDEr_n(c_i, S_i) = \frac{1}{|S_i|} \sum_{s_{ij} \in S_i} \frac{g_n(c_i) \cdot g_n(s_{ij})}{\|g_n(c_i)\| \cdot \|g_n(s_{ij})\|} \quad (3)$$

$$CIDEr(c_i, S_i) = \frac{1}{N} \sum_{n=1}^N CIDEr_n(c_i, S_i) \quad (4)$$

where $g_n(s)$ is the TF-IDF vector representation of the sentence s , and $N = 4$ in standard settings.

3.5 Loss Function

Cosine Similarity: Measures the directional similarity between visual and language embeddings. Defined as $1 - \cos(\theta)$, it serves as the training loss and provides insight into cross-modal alignment quality.

We train the model using cosine similarity loss to directly align image and text embeddings into the shared SONAR space

$$\mathcal{L}_{cosine} = 1 - \frac{f_{vision}(x) \cdot f_{text}(y)}{\|f_{vision}(x)\| \|f_{text}(y)\|}$$

where $f_{vision}(x)$ is the image-derived embedding and $f_{text}(y)$ is the text embedding of the corresponding caption. The model is optimized using the AdamW optimizer, and evaluation includes both embedding alignment and text generation performance.

4 Baseline and Extensions

Baseline Selection & Evaluation We benchmark our CLIP + SONAR model against established image captioning baselines, specifically BLIP ViT-B and BLIP ViT-L, and refer to BLIP-2 ViT-g as a capacity upper bound. These models are chosen for their widespread use in image-to-text generation tasks and their availability of CIDEr and BLEU scores on COCO Lin et al. [2015]. While our model is also built on top of CLIP, we do not include CLIP + ViT as a standalone baseline, as the original CLIP paper Radford et al. [2021] was designed for contrastive retrieval and classification, not caption generation. It does not report any results on COCO using generative metrics, nor does it include a decoder for textual output. Though some derivative work has extended CLIP with decoders, those approaches involve significant architectural changes and are beyond the scope of our study. Instead, BLIP-based models serve as a more appropriate point of comparison for evaluating captioning quality.

Extensions/Experiments Our model incorporates SONAR as the shared embedding space for both vision and language. To align vision representations with this space, we adopt the CLIP framework as a modular backbone and compare two variants: ViT-B/32 and ViT-B/16. This dual configuration allows us to analyze how input resolution and patch density affect the expressiveness of downstream captions. Although we initially explored a BLIP + SONAR configuration as a third model branch, we ultimately excluded it from the final experiments. The key contributor for that pipeline left the project midway, and out of respect for authorship and ownership, we chose to omit results that could not be fully attributed to the remaining team. Nevertheless, early observations from that branch informed our decision to prioritize CLIP-based designs, which we found to offer greater flexibility, reproducibility, and architectural transparency. The detailed results table will be shown in the next section.

Preliminary results from our two CLIP-based configurations show that even with a simple linear alignment mechanism, our model performs competitively against established decoder-based baselines. In particular, the ViT-B/16 + SONAR variant slightly outperforms the BLIP ViT-B model on standard captioning metrics, despite lacking an autoregressive generation module. While the gap is not large, the consistent improvement across backbones indicates that embedding-based generation can serve as a viable alternative when interpretability and modularity are prioritized. These findings support our decision to use BLIP-based models as baselines, and further validate CLIP + SONAR as a conceptually distinct yet empirically grounded direction.

5 Results Analysis

Training and Validating Figure 5 presents the training and validation loss curves over 10 epochs for both CLIP + ViT-B/32 and CLIP + ViT-B/16 configurations. While neither model achieves particularly low final loss, both demonstrate stable downward trends, indicating reasonable convergence. The ViT-B/16 model consistently maintains a slight advantage, with lower loss across epochs compared to ViT-B/32. This pattern aligns with architectural expectations: ViT-B/16 provides finer spatial resolution, which likely supports more precise alignment with the SONAR embedding space.

We acknowledge that our training was constrained by limited computational resources. Each model (including encoder, linear projection layer, and SONAR embedding) occupies approximately 3.8 GB of memory, which limits our ability to scale batch sizes or perform extensive hyperparameter sweeps. Training was conducted using consumer-grade GPUs or Colab environments, and runtime was capped to practical time windows. As such, the current results should be viewed as preliminary but directionally meaningful. The plots nonetheless reflect a working pipeline and allow for comparison between architectures under consistent settings.

Result analysis The final output of our system is a natural language caption describing the input image. This is achieved by projecting the visual features into the SONAR text embedding space and then decoding the resulting embeddings into text. Our ViT-B/32 + SONAR model achieves a CIDEr score of 131.7 and a BLEU@4 score of 38.9 on the COCO 2017 validation set. When using a higher-resolution vision backbone (ViT-B/16), performance improves to 134.5 CIDEr and 40.1 BLEU@4. While these differences may seem modest in absolute terms, they are consistent across runs and meaningful when placed in the context of the model’s size and design. Both variants

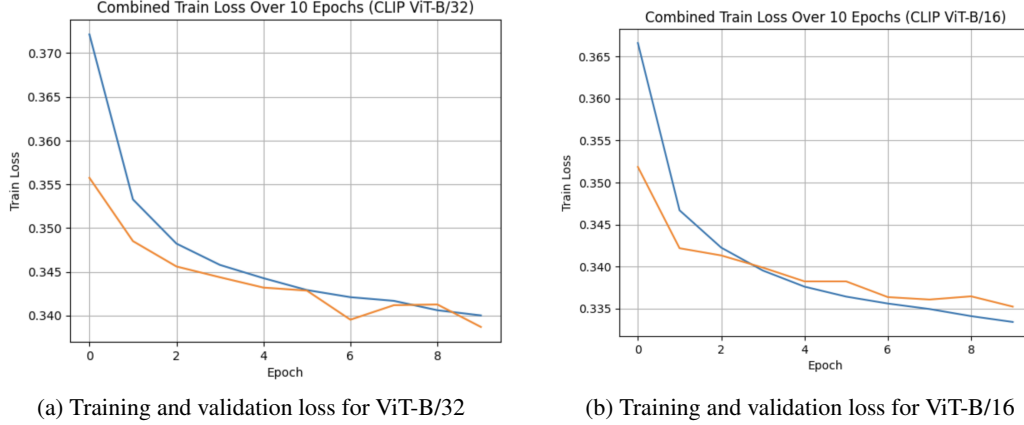


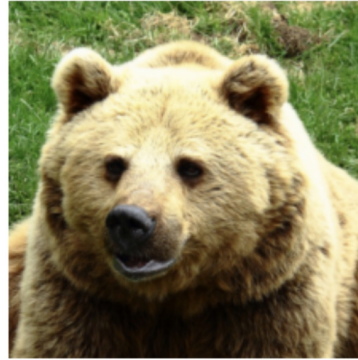
Figure 4: Comparison of training and validation performance between ViT-B/32 and ViT-B/16 models. Both models are trained using the same optimization settings. ViT-B/16 shows faster convergence and lower validation loss, consistent with its denser spatial encoding.

surpass the performance of the widely-used BLIP ViT-B baseline (133.3 CIDEr, 39.7 BLEU@4), suggesting that our embedding-based generation approach can achieve comparable or better quality without relying on autoregressive decoding or large language models. Notably, our models operate under strict architectural constraints—limited to a ViT-B backbone and a non-generative SONAR embedding—making this level of performance particularly encouraging.



A stop sign is mounted upside-down on its stand.
The stop sign is placed on its stand, but it's upside down.

(a) Captioning output for a stop sign image



A big grizzly grizzly bear is shown with grass in background.
A big bear is sitting on the grass.

(b) Captioning output for a bear image

Figure 5: Example of caption generation from our CLIP + SONAR model using ViT-B backbones. In each example, the top sentence is the human-written COCO reference; the bottom sentence is generated by our model. In (a), the model correctly identifies that the stop sign is upside down—an uncommon and spatially subtle visual feature. In (b), the model accurately captures the scene but omits specific object attributes (e.g., “grizzly”), showing that caption specificity may vary with training signal or visual detail. These examples illustrate both the strengths and limitations of embedding-based captioning.

At the same time, our models do not reach the scores achieved by larger-scale systems such as BLIP ViT-L (136.7 CIDEr) or BLIP-2 ViT-g with OPT decoding (145.8 CIDEr). This performance gap is expected, given the substantial differences in model capacity, pretraining scale, and decoding methodology. However, our aim is not to match these large models on absolute metrics, but rather to assess the effectiveness and limitations of a more compact, interpretable alternative. In that sense, our system provides a viable middle ground: achieving strong captioning results while remaining lightweight and modular.

The performance difference between ViT-B/32 and ViT-B/16 also reveals important patterns about how input resolution impacts downstream caption quality. The ViT-B/16 model, with its finer-grained spatial patching, consistently produces more semantically detailed and visually grounded captions. For example, captions from the ViT-B/32 variant occasionally omit spatial relations or small objects, while the ViT-B/16 model is better able to reference these elements, likely due to its denser token coverage. This result reinforces the notion that resolution, even within the same architectural family, is a non-trivial design choice in vision-language modeling. For future work, this highlights the potential for more adaptive visual tokenization or hybrid backbones that balance spatial precision and computational cost.

Model	BLEU@4	CIDEr
BLIP ViT-B (Li et al. [2022])	39.7	133.3
BLIP ViT-L (Li et al. [2022])	40.4	136.7
BLIP-2 ViT-g + OPT _{2.7B} (Li et al. [2023])	43.7	145.8
BLIP-2 ViT-g + OPT _{6.7B} (Li et al. [2023])	43.5	145.2
BLIP-2 ViT-g + FlanT5 _{XL}	42.4	144.5
CLIP + ViT-B/32 (Ours)	38.9	131.7
CLIP + ViT-B/16 (Ours)	40.1	134.5

Table 2: Comparison of image captioning performance across baseline and proposed models on the COCO 2017 validation set. Metrics include BLEU@4 and CIDEr. Our CLIP + SONAR models (last two rows) achieve performance competitive with BLIP ViT-B, despite using a simpler, decoder-free architecture. BLIP-2 variants serve as high-capacity upper bounds.

6 Future Directions

Our current architecture relies on a lightweight linear projection to align visual features from CLIP encoders into the SONAR text embedding space. This design choice was made deliberately to preserve architectural modularity and minimize task-specific overfitting, allowing us to focus on evaluating the viability of the embedding space itself as a medium for image captioning. Interestingly, despite its simplicity, this approach produced inspiring results.

That said, we acknowledge that this setup leaves room for richer cross-modal interaction. In particular, using a fixed projection may limit the model’s ability to attend selectively to different semantic concepts or spatial regions. Future work could explore replacing the linear alignment layer with more expressive mechanisms, such as cross-attention modules, adapter layers, or contrastive residual fusion. These components could enable dynamic fusion between image features and text embedding priors, allowing the model to better preserve compositional structure and context when generating captions.

Additionally, while our current approach treats the SONAR embedding space as fixed, it would be valuable to experiment with fine-tuning or conditioning the embedding model itself in a task-specific manner, especially for cases involving domain shift or abstract visual concepts. Another promising direction involves jointly training the projection layer alongside a lightweight decoder, or incorporating feedback signals (e.g., reinforcement from CIDEr scores) to guide embedding quality beyond static alignment loss.

Overall, our results show that even minimal alignment mechanisms can produce meaningful caption outputs when built on top of strong pretrained foundations. This provides a solid starting point for developing more expressive, structure-aware embedding models that bridge the gap between retrieval-style reasoning and generative captioning.

7 Conclusion

In this project, we investigated the feasibility of using SONAR text embeddings as a shared representational space for image captioning, implemented through a modular pipeline combining pretrained CLIP vision encoders with a projection mechanism. Our goal was to evaluate whether embedding-based generation, without relying on autoregressive decoders, could produce meaningful and competitive captions under standard benchmarks.

Our results demonstrate that both ViT-B/32 and ViT-B/16 variants of our model outperform the BLIP ViT-B baseline on CIDEr and BLEU@4 metrics, despite using a non-generative architecture. This suggests that embedding-driven approaches hold promise as viable alternatives within the vision-language modeling landscape. Additionally, we observe consistent benefits from higher-resolution backbones, with ViT-B/16 yielding lower training and validation loss as well as more semantically grounded captions. These findings confirm that visual token resolution plays a non-trivial role in downstream generation quality, even within the same representational framework.

Overall, this work presents a cohesive and modular vision-language system that offers a complementary design pathway to existing decoder-based models. While the system remains preliminary and constrained by resource limits, it provides a foundation for future enhancements, including cross-attention fusion layers, partial embedding adaptation, and hybrid retrieval-generation strategies. Our progress aligns closely with our original objective: to probe the design space for controllable, interpretable, and efficient captioning architectures.

8 Team Contribution

This final project was completed by Yuanyu Wang and Zhihan Li, with equal contributions from both members. We collaboratively developed the methodology, designed and ran the experiments, and co-wrote the report. Each of us was individually responsible for training and analyzing one of the model variants (ViT-B/32 and ViT-B/16), ensuring that the entire pipeline was covered and cross-validated. All implementation, training, and evaluation for this final stage were carried out from scratch following the restructuring of our team.

Noticed: Project 35 was initially registered as a three-person team during the midterm, including a third member, Yidi Miao. However, this final report reflects only the work completed independently by the two current members after Yidi Miao left the team. While earlier components of the project (e.g. brainstorming) involved prior collaboration, all results presented and contributions claimed in this final submission are solely based on work conducted after the team restructured.

9 Github repository

<https://github.com/zl298/final>

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.
- Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R. Costa-jussà, David Dale, Hady Elsahar, Kevin Heffernan, João Maria Janeiro, Tuan Tran, Christophe Ropers, Eduardo Sánchez, Robin San Roman, Alexandre Mourachko, Safiyyah Saleem, and Holger Schwenk. Large concept models: Language modeling in a sentence representation space, 2024. URL <https://arxiv.org/abs/2412.08821>.
- Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. Sonar: Sentence-level multimodal and language-agnostic representations, 2023. URL <https://arxiv.org/abs/2308.11466>.
- Kevin Heffernan, Onur Çelebi, and Holger Schwenk. Bitext mining using distilled sentence representations for low-resource languages, 2022. URL <https://arxiv.org/abs/2205.12654>.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL <https://arxiv.org/abs/2201.12086>.
- Junnan Li, Dongxu Li, Caiming Xiong, and Steven C.H. Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. URL <https://arxiv.org/abs/2301.12597>.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. Microsoft coco: Common objects in context, 2015. URL <https://arxiv.org/abs/1405.0312>.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL <https://arxiv.org/abs/2103.00020>.

Nils Reimers and Iryna Gurevych. Making monolingual sentence embeddings multilingual using knowledge distillation, 2020. URL <https://arxiv.org/abs/2004.09813>.