Vision-Language Pretraining for Image-to-Text Generation

Pranshul Bhatnagar, Aesha Gandhi, Gaurav Law

1 Introduction

Recent advances in Vision-Language Pretraining (VLP) have greatly improved bidirectional understanding between visual and textual modalities. Models such as Bootstrapped Language Image Pretraining (BLIP) and Vision Transformer (ViT) now excel at tasks such as visual question answering (VQA) and image retrieval. However, while text-to-image generation has received significant attention, the reverse task — generating rich, context-aware text from images — remains less explored and more challenging.

This project investigates whether VLP models can effectively invert this relationship by producing accurate, semantically meaningful descriptions from images. We focus on ViT and BLIP due to their central role in current text-to-image pipelines. ViT serves as a pure vision encoder, whereas BLIP integrates both image and text inputs within a multimodal training loop. BLIP's architecture wraps ViT within a framework that learns joint representations, making it well-suited for our reverse-generation task.

Using pre-collected image–prompt pairs, we train models to generate new captions, then evaluate them through validation and testing. Performance is assessed using established metrics such as BLEU and BERTScore to measure linguistic quality and semantic alignment.

2 Data

For the purpose of this project, we sourced our data from DiffusionDB, a large-scale text-to-image dataset that contains over 14 million images generated with Stable Diffusion 2.0 using user-provided prompts and hyperparameters. From this collection, we randomly selected approximately 10,000 images, each with a resolution of 512×512 pixels, along with their corresponding English prompts, which varied widely in length (distribution shown in Figure 1). Random sampling allowed us to capture a broad and unbiased subset of the dataset while keeping the computational load manageable. Although more targeted sampling strategies could have revealed specific trends, using a random selection helped ensure that our results remained generalizable across a wide range of prompt-image pairs.

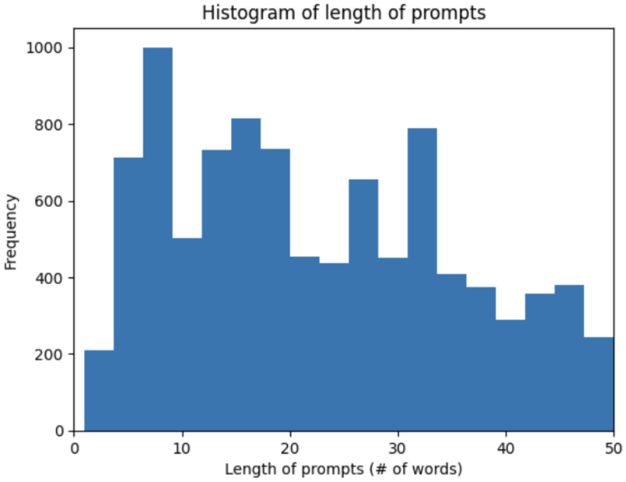


Figure 1: Distribution of Length of prompts in our Dataset

Following this, the dataset was split into train (60%), validation (20%), and test (20%) sets using a fixed seed of 42 to ensure reproducibility. The training set was used to teach the models to learn patterns between images and their associated prompts. The validation set was used to monitor the model's performance during training, allowing us to fine tune hyperparameters and prevent overfitting. Finally, the testing set was used to evaluate the model's performance on unseen data, ensuring an unbiased assessment of generalization.

3 Model Architecture

3.1 Vision Transformer + GPT-2

The Vision Transformer applies the transformer architecture by splitting images into small patches (such as 16 × 16 pixels), flattening them, and projecting into embeddings so each patch becomes a "token." Positional embeddings preserve spatial structure, and the sequence of patch tokens is processed by a transformer encoder (12 transformer blocks) whose self-attention layers build a global understanding of the image. The resulting visual embedding is fed into the GPT-2 decoder (12 transformer blocks with cross-attention), which generates the caption one word at a time as an autoregressive model, forming the encoder–decoder setup. Inference is performed using beam search with four beams and early stopping.

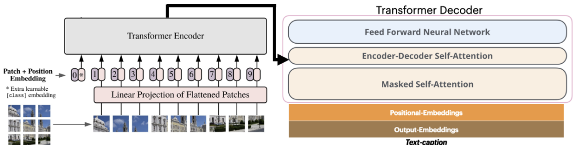

Transformer Encoder
Patch + Position Embedding
* Extra learnable [class] embedding
0
*
1
2
3
4
5
6
7
8
9
Linear Projection of Flattened Patches
Transformer Decoder
Feed Forward Neural Network
Encoder-Decoder Self-Attention
Masked Self-Attention
Positional-Embeddings
Output-Embeddings
Text-caption

Figure 2: ViT-GPT2 Model Architecture

3.2 Bootstrapping Language-Image Pre-Training

Bootstrapping Language-Image Pre-Training is a generative method using an encoder-decoder architecture enhanced by a caption-filtering mechanism (Li et al., 2022). BLIP is designed to generate accurate image captions while filtering out noisy image-text pairs, allowing it to effectively leverage large-scale pre-training data. Its Multimodal Encoder-Decoder (MED) model consists of: (1) Unimodal encoder trained with image-text contrastive (ITC) loss to align visual and textual representations, (2) Image-grounded text encoder which incorporates visual information through cross-attention layers and is trained with image-text matching (ITM) loss to distinguish correct from incorrect image-text pairs, and (3) Image-grounded text decoder that generates captions using a language modeling (LM) loss, employing causal self-attention along with cross-attention layers to capture interactions between vision and language.

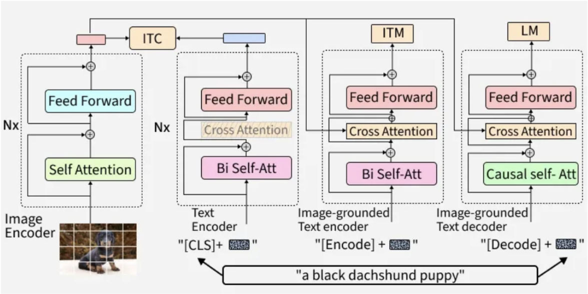


***Fig 3:** BLIP Model Architecture*

Figure 3: BLIP Model Architecture

BLIP has shown strong performance in both vision-language understanding and generation tasks; because of its robust pre-training approach and proven results, we selected BLIP as one of the candidates for our experimentation.

4 Methodology

Modern image-to-text systems typically rely on neural encoder–decoder architectures, such as transformer-based models, which excel at generating coherent text sequences from visual inputs. Because our images were produced through a similar generative process, leveraging transformers for caption generation offers a natural alignment between the image representations and the text decoding pipeline.

To strengthen the model's performance, we incorporated both pre-training — enabling the model to learn broad visual and linguistic patterns — and transfer learning, allowing us to adapt those learned features to our specific dataset with far fewer resources. Together, these techniques improve accuracy, reduce training time, and increase overall effectiveness in text generation.

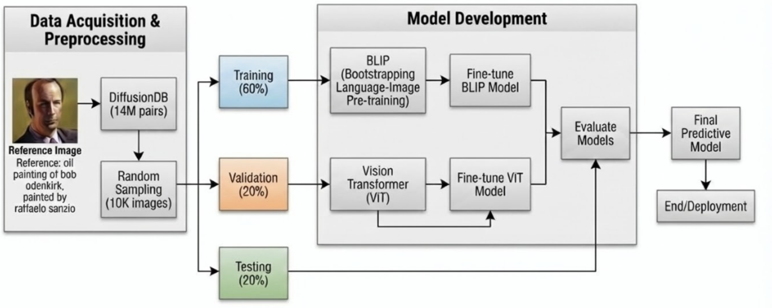


Figure 4: Flowchart outlining our Methodology

We implemented a PyTorch Dataset class to handle loading and preprocessing, converting images into tensors and captions into tokenized sequences using the appropriate processor. The dataset was split into training and validation sets, shuffled to ensure randomness, and loaded in batches to optimize GPU memory usage. Training for both models was performed on a NVIDIA A100 GPU using a sequence-to-sequence objective with cross-entropy loss.

5 Evaluation Metric

To evaluate the performance of both the BLIP-and-ViT-based captioning models, we developed a custom metric that combines Bilingual Evaluation Understudy (BLEU) and Bidirectional Encoder Representations (BERT) Score to capture semantic meaning and lexical accuracy. BLEU measures n-gram overlap between the generated caption and the reference caption, which directly evaluates the exactness of the output. BERT Score computes similarity using contextual embeddings from a transformer encoder, allowing it to capture deeper semantic alignment even when the wording differs.

To emphasize semantic correctness without ignoring lexical precision, we assigned 65% weight to BERTScore and 35% weight to BLEU:

$$\text{Score} = 0.65 \times \text{BERTScore} + 0.35 \times \text{BLEU} \quad (1)$$

All metrics were computed on cleaned predictions for consistency. Both the pretrained baseline and the fine-tuned model used the same inference pipeline.

6 Results & Discussions

6.1 BLIP

For fine-tuning BLIP, we used the pre-trained model Salesforce/blip-image-captioning-base and trained it for 10 epochs. Fine-tuning allowed the model to adjust its pre-learned visual and linguistic representations to better align with the specific distribution and style of our dataset. We employed a learning rate of 1e-4 with a CosineAnnealingWarmRestarts scheduler, which periodically increases the learning rate to help the model escape local minima. A weight decay of 1e-6 was used, and gradient accumulation over 4 steps allowed for effective batch size scaling. The maximum sequence length for captions was set to 128 tokens.

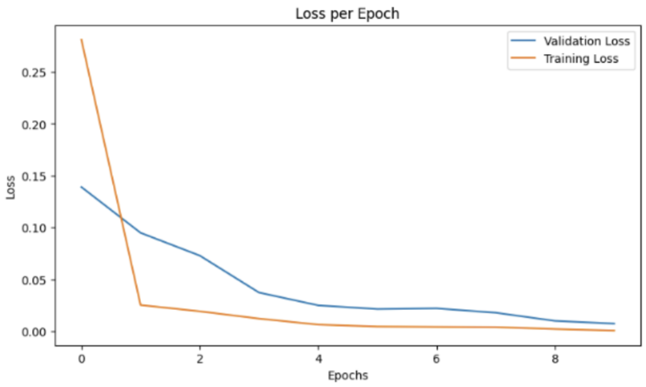


Figure 5: Training vs Validation Loss (BLIP)

During fine-tuning, both training and validation losses decreased steadily, reflecting the model's ability to learn from the dataset while generalizing well to unseen samples. As shown in Figure 5, the validation loss consistently followed the trend of the training loss, and the lowest validation loss was achieved at the final epoch, indicating stable and effective fine-tuning without evidence of overfitting. The fine-tuned BLIP model achieved a lowest validation loss of 0.0075 by the end of the 10th epoch.

Table 1: BLIP Fine-tuning Results

Metric	Pre	Fine	Change	Improv
BERT	0.44	0.76	+0.32	+73%
BLEU	0.27	0.78	+0.51	+189%
Comb.	0.38	0.77	+0.39	+102%

The fine-tuned BLIP model shows substantial improvements over the pretrained baseline as shown in Table 1. BERTScore increased by over 73%, indicating that the fine-tuned model produced captions that were much more contextually aligned with the reference text. BLEU improved the most, rising by nearly 189%, showing that the model learned to generate more precise and textually accurate descriptions. When combined, the fine-tuned model achieved an improvement of about 102%, demonstrating that fine-tuning significantly enhanced both the semantic and lexical quality of the generated captions.

When we looked at BLIP's predictions manually (Table 3), there's a clear difference between the captions it handled well and the ones it struggled with. For the good predictions, the model did a strong job capturing both the overall theme of the image and small details, especially when the image had a clear subject, a distinct art style, or strong visual cues. The bad predictions show some of BLIP's weaknesses; in images that were more ambiguous or had fewer visual anchors, the model tended to guess or hallucinate details that were not actually present.

6.2 ViT

Our dataset images were converted to RGB and processed using the ViTImageProcessor, which handled normalization and resizing to the pretrained ViT encoder's expected input size. Captions were tokenized using the GPT-2 tokenizer with a maximum length of 80 tokens; padding tokens were replaced with -100 placeholders to exclude them from loss computation.

Fine-tuning used the pretrained nlpconnect/vit-gpt2-image-captioning VisionEncoderDecoderModel. Training was conducted for six epochs with a per-device batch size of four and gradient accumulation over two steps for an effective batch size of eight. The AdamW optimizer was used with a learning rate of 5×10^{-5}, and mixed-precision FP16 training was enabled for memory efficiency on GPU.

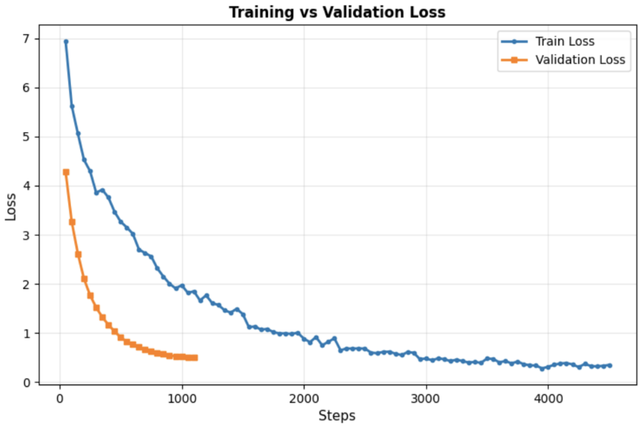


Figure 6: Training vs Validation Loss (ViT-GPT 2)

As shown in Figure 6, training loss declined smoothly, from ∼6.9 at initialization to below 1.0 by epoch 6, and validation loss followed the same downward trend, indicating successful learning and stable generalization.

Table 2: ViT GPT2 Fine-tuning Results

Metric	Pre	Fine	Change	Improv
BERT	0.42	0.43	+0.01	+3%
BLEU	0.00	0.00	+0.00	+163%
Comb.	0.27	0.28	+0.01	+3%

As shown in Table 2, the fine-tuned ViT captioning model shows small improvements over the pretrained baseline. While the absolute values remain modest, the consistent upward movement across metrics shows that the ViT model did benefit from task-specific training and became better aligned with the captioning style of the dataset. However, a qualitative review of the example outputs (Table 4) shows that the model still struggles with generating detailed and fully accurate descriptions, occasionally generating repetitive or incomplete phrases.

We also noticed that images with higher BERTScores tended to produce captions that were more coherent and better aligned with the intent of the prompt, even in cases where the combined score was more moderate. This suggests that BERTScore captures semantic alignment particularly well in our setting and better reflected whether a caption made sense to a human reader, even when the wording differed from the ground truth.

7 Conclusion

This project investigated whether the text-to-image process can be meaningfully reversed by generating detailed, context-aware descriptions from images. Comparing a ViT-GPT encoder–decoder model with the multimodally pre-trained BLIP model revealed a clear advantage for models grounded in joint vision–language pre-training. BLIP showed substantial gains after fine-tuning, capturing objects, styles, and visual details more effectively, whereas the ViT-GPT model achieved only modest improvements due to its limited semantic grounding.

Despite BLIP's strong performance, both models struggled with deeper contextual reasoning, abstract concepts, and complex relationships, often producing captions shorter and less expressive than target prompts. These limitations highlight the need for richer training data, improved architectures, and more semantically sensitive evaluation metrics.

Overall, our findings demonstrate that reversing the text-to-image relationship is achievable but highly dependent on the model's pre-training paradigm. Advancing this direction through larger datasets, better tokenization, refinement techniques, and architectures built for bidirectional vision language tasks holds promising potential for applications in accessibility, robotics, retrieval, and multimodal AI systems.

Table 3: Example Prompt Predictions using our Fine-tuned BLIP Model

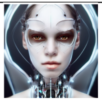
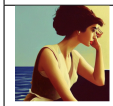
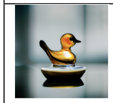
Image	Prompt	Caption	BLEU	BERT	Score
GOOD PREDICTIONS					
	A portrait of a beautiful female cyborg with a cracked porcelain face by wlop, exposed inner structure, brightly glowing white eyes, art nouveau card, trending on artstation	a portrait of a beautiful female cyborg wearing an intricate venetian mask by Wlop	0.79	0.69	0.73
	frontal portrait of a muscular woman in a long - sleeved linen shirt, wearing vintage engineer goggles, by norman rockwell	a portrait of a muscular anthropomorphic cyberpunk iguana! in leather spacesuit	0.51	0.58	0.52
	top-heavy 20 year old with messy black hair and big beard eats mayonnaise straight out of the jar with his bare hands	obese man with messy black hair and big beard drinking mayonnaise from the jar	0.64	0.76	0.72
BAD PREDICTIONS					
	jack vettriano painting of lilly collins	a beautiful british woman with short brown hair, gentle, vibrant amber eyes, standing on a rooftop	0.17	0.22	0.1999
	A house with a beautiful backyard.	photorealistic beautiful cherry blossom forest with victorian lanterns lining the stone pathway. hyperdetailed photo	0.01	0.51	0.33

Table 4: Example Prompt Predictions using our Fine-tuned ViT-GPT2 Model

Image	Prompt	Caption	BLEU	BERT	Score
GOOD PREDICTIONS					
	a close up photo of a glass duck on a table, visible background, professional photography	close photo a duck a of glass on table visible, background professional	0.43	0.72	0.48
	photo still of al roker in a ball pit!!!!!!! at age 4 6 years old 4 6 years of age!!!!!!!	still christ a loft !!!!!!!! age 5 old fun pop by bag bund !!!!!!!! age 5 old 5 of!!!!!!!	0.09	0.64	0.44
	A portrait of a beautiful anime girl wearing a long coat, shoulder length straight black hair	cute girl a anime wearing stylish coat shoulder straight, length bob hair straight hair	0.01	0.66	0.43
BAD PREDICTIONS					
	Alstom tr train on track	blue lights	0.00	0.20	0.13
	amazing futuristic design for a 3D printer, intricate, highly detailed, smooth, sharp focus	of all	0.01	0.20	0.13

References

1. A. Kumar, "The Illustrated Image Captioning using transformers," Ankur — NLP Enthusiast, Nov. 20,2022.https://ankur3107.github.io/blogs/the-illustrated-image-captioning-using-transfo rmers/ (accessed Mar. 08, 2023)

2. Li, J., Li, D., Xiong, C., & Hoi, S. (2022). Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In International Conference on Machine Learning (pp. 12888-12900). PMLR.

3. Ordonez, V., Kulkarni, G., & Berg, T. (2011). Im2text: Describing images using 1 million captioned photographs. Advances in neural information processing systems, 24.

4. Radford A, Kim JW, Hallacy C, et al. Learning transferable visual models from Natural Language Supervision. arXiv.org. https://arxiv.org/abs/2103.00020. Published February 26, 2021. Retrieved April 5, 2023.

5. Reimers N, Gurevych I. Sentence-bert: Sentence embeddings using Siamese Bert-Networks. arXiv.org. https://arxiv.org/abs/1908.10084. Published August 27, 2019. Accessed April 12, 2023.

6. Singhal, A. (2001). Modern information retrieval: A brief overview. IEEE Data Eng. Bull., 24(4), 35-43.

7. Wikimedia Foundation. (2023, February 26). Cosine similarity. Wikipedia. Retrieved April 5, 2023, from https://en.wikipedia.org/wiki/Cosine_similarity

8. Zhu Y, Lu S, Zheng L, et al. Texygen: A benchmarking platform for text generation models. arXiv.org. https://arxiv.org/abs/1802.01886. Published February 6, 2018. Accessed April 15, 2023.