# A Unified Model for Image-Text Semantic Alignment and Prompt Generation Based on Multi-Task Collaborative Training

**Authors:** Qilong Du 4743340, Junfeng Wei 4742674

**Code Repository:**
https://github.com/QilongDuTony/GNN-final-project.git

**Lecture:** Generative Neural Network

**Date:** March 10, 2025

# Contents

# 1 Abstract

**Author: Qilong Du 4743340, Junfeng Wei 4742674 both contributed equal teamwork**

The goal of this study is to generate meaningful and semantically rich textual prompts from images, a task that plays a pivotal role in bridging the gap between visual understanding and natural language generation. A crucial prerequisite for this task is ensuring robust image-text alignment, which directly affects the quality of the training data and the overall performance of prompt generation models. Poorly aligned image-text pairs can mislead the model and result in semantically inconsistent outputs. Therefore, we first focus on solving the image-text alignment problem to establish a strong foundation for the downstream image-to-prompt generation task.

To this end, we propose a novel framework that utilizes a pre-trained ResNet50 as the image encoder and BERT as the text encoder. A fully connected layer is employed to project both image and text embeddings into a shared semantic space. The model is trained using cosine similarity loss, which encourages the embeddings of semantically related image-text pairs to be close in the latent space. This alignment ensures that the representations of both modalities capture similar semantic information, which is vital for subsequent prompt generation tasks.

We conduct experiments on the Stable Diffusion - Image to Prompts dataset available randomly on HuggingFace. By computing the cosine similarity between seven typical extracted image embeddings and BERT-encoded prompts, we assess the model's alignment performance. Our results show consistently high similarity scores ranging from 0.77 to 0.94, indicating that our approach effectively captures the semantic correspondence between modalities.

Our contributions are twofold: first, we introduce an effective method for aligning visual and textual representations, which can be applied to any multimodal task; second, we highlight the importance of robust image-text alignment as a foundational step for training reliable image-to-prompt models. This work advances the field of multimodal learning and provides a solid basis for future research in cross-modal semantic understanding.

Building on this alignment framework, we explore a second stage of our research: training a generative model that can produce prompts directly from images.

We employ a pre-trained ResNet50 as the image encoder and GPT-2 as the language decoder. The model is trained on a subset of the DiffusionDB dataset, which provides a large-scale pairing of synthetic images and their corresponding prompts.

Our goal is to evaluate whether the encoder-decoder model can effectively learn to generate descriptive and accurate prompts that resemble the original inputs used for image generation. We assess the model's performance using qualitative comparisons and loss analysis across multiple epochs.

Through this project, we contribute a novel approach to prompt generation, and insights into the challenges and limitations of current methods. This work

also sets the foundation for future research in automated content annotation for specific tasks, multimodal retrieval systems, and closed-loop generative learning.

## 2 Introduction

**Author: Qilong Du 4743340, Junfeng Wei 4742674 both contributed equal teamwork**

In recent years, the field of Vision-and-Language (V&L) learning has gained significant momentum, driven by advances in both natural language processing (NLP) and computer vision (CV). A prominent subdomain within this field is the task of text-to-image generation, which aims to generate realistic, diverse, and semantically coherent images from textual descriptions. This subfield has wide-ranging applications, including but not limited to digital content creation, design automation, education, virtual environments, and accessibility support for visually impaired users. The overarching goal of this research area is to bridge the semantic gap between textual and visual modalities, enabling machines to comprehend, reason about, and generate cross-modal content in a human-like manner.

The first part of the project focuses on the reverse mapping: given an image, can we reconstruct or generate a descriptive and semantically meaningful prompt? This inverse task is particularly interesting because it challenges the model to learn rich cross-modal associations. If successful, such models could be useful for automatic content labeling, improving accessibility, or serving as components in interactive generative systems.

The context of this research lies at the intersection of computer vision, natural language processing (NLP), and deep learning. More specifically, it resides within the field of image captioning and cross-modal representation learning. While image captioning typically involves generating a brief sentence that describes the visual content (e.g., "A cat sitting on a couch"), our task is subtly but critically different. Instead of describing a scene, the goal is to reproduce a prompt that could have originally generated the image in a generative model pipeline. These prompts often include stylistic cues, scene composition, and artistic elements—thus requiring deeper semantic understanding.

To address this challenge, we investigate a transformer-based encoder-decoder architecture. The encoder is a ResNet50 convolutional neural network pretrained on ImageNet, responsible for extracting image features. The decoder is a GPT-2 language model that generates the corresponding prompt token by token, autoregressively. This architecture aligns with prior work in neural machine translation and image captioning but adapts the approach to a different form of supervision—namely, text prompts that guided synthetic image generation.

The dataset used in this study is a curated subset of DiffusionDB, which contains over two million image-prompt pairs generated by Stable Diffusion. From this large corpus, we randomly selected 1000 samples to fine-tune our encoder-decoder pipeline. The diversity and richness of these prompts make the dataset suitable for studying both semantic understanding and stylistic generation.

The motivation behind this part of research is multi-faceted. First, there is a growing demand for automatic image annotation tools that can generate human-like descriptions of content, particularly in scenarios where metadata is missing. Second, the ability to reconstruct prompts enables deeper inspection and interpretability of generative models. Third, prompt generation has applications in creative industries, recommendation engines, and even in the training of prompt-based agents.

To effectively generate prompts that are both relevant and semantically aligned with the input images, it is essential to first ensure that the training data itself reflects strong image-text correspondence. While the encoder-decoder model forms the core of our generation task, our research also incorporates a complementary objective: validating and improving image-text alignment quality. This ensures that the data used for generation truly captures the intended semantics. To achieve this, we develop a separate but related alignment model using ResNet50 and BERT, trained with cosine similarity loss in a shared embedding space. Together, these two components—generation and alignment—form a complete pipeline for reliable prompt synthesis from images. One of the most impactful innovations in this space is the development of diffusion models [6], which have emerged as powerful generative frameworks capable of producing high-fidelity images from text prompts. These models leverage iterative denoising techniques to generate images from noise while being guided by the semantics of the input text. By integrating multimodal input mechanisms, diffusion models like Stable Diffusion excel at generating visually compelling and diverse outputs. Their open-source nature has also accelerated research and development across academia and industry. However, despite these advancements, one of the persistent challenges in diffusion-based models is achieving precise alignment between the generated images and their corresponding textual prompts.

This alignment issue becomes especially prominent when dealing with complex or compositional prompts. For instance, prompts describing multiple objects, detailed attributes, or abstract relationships often result in partial or misaligned visual outputs. These inconsistencies stem not only from limitations in the generation model but also from noise and ambiguity in the training data. A significant portion of text-to-image datasets contains loosely or inaccurately paired image-text pairs, which weakens the model's ability to learn robust cross-modal mappings. To tackle these challenges, researchers have proposed optimization strategies such as Fast Prompt Alignment (FPA) [5], which aim to refine prompt formulations for better alignment during image synthesis. Nevertheless, such strategies frequently assume that the underlying image-text pairs are already semantically accurate, an assumption that does not hold true in many real-world scenarios.

Given this context, we argue that effective text-to-image generation must begin with a foundational step: robust image-to-text alignment. Before attempting to generate prompts or train generative models, it is crucial to ensure that image-text pairs used for training are semantically consistent. This is especially important for tasks like reverse prompt generation, prompt tuning, and training interpretable multimodal models. Without strong alignment at the

data level, downstream models risk learning spurious correlations and producing unreliable outputs. Therefore, we focus on solving the image-text alignment problem as a prerequisite for reliable prompt generation.

In this study, we propose a novel framework for measuring and enhancing image-to-text alignment by mapping both images and textual prompts into a shared semantic space. To achieve this, we adopt a pre-trained ResNet50 [3] as the image encoder and BERT [1] as the text encoder. ResNet50 is a well-established convolutional neural network known for its high performance in image classification and feature extraction. It efficiently captures hierarchical visual patterns and represents images as dense feature embeddings. On the textual side, BERT (Bidirectional Encoder Representations from Transformers) is a pre-trained language model that transforms natural language into context-aware semantic embeddings. By utilizing BERT, we ensure that complex sentence structures and prompt semantics are effectively captured.

To align the two modalities, we introduce a fully connected projection layer that maps both image and text embeddings into a shared latent space. We then apply cosine similarity loss during training to ensure that semantically similar image-text pairs are close in this space. This method does not require explicit supervision, making it scalable and adaptable to a wide range of datasets. Our design choice to use cosine similarity is based on its effectiveness in preserving semantic relationships in high-dimensional embedding spaces, although we acknowledge that it may not capture all forms of semantic nuance.

We evaluate our method using a subset of the Stable Diffusion - Image to Prompts dataset, randomly selected from HuggingFace. We compute the cosine similarity between image embeddings generated by ResNet50 and text embeddings generated by BERT across seven representative image-text pairs. Our experimental results show high alignment quality, with similarity scores ranging from 0.77 to 0.94. These findings indicate that our model effectively captures the underlying semantics of both images and text, and confirms the feasibility of our alignment strategy.

Our contributions in this paper are threefold. First, we emphasize the importance of robust image-text alignment as a critical step for successful prompt generation. Second, we introduce a reproducible and lightweight model architecture for embedding alignment using widely available pre-trained models. Third, we provide empirical evidence that high-quality alignment improves dataset reliability and supports better training for downstream image-to-prompt tasks.

However, our approach is not without limitations. Since we rely on pre-trained encoders, the embeddings may not be perfectly optimized for multimodal alignment without fine-tuning. Additionally, cosine similarity may not fully capture visual scene structure or complex spatial relationships. In future work, we plan to explore attention-based mechanisms and multimodal contrastive learning to further refine alignment quality.

In summary, this work lays a foundation for building better-aligned image-text datasets, which is essential for training accurate and interpretable prompt generation systems. We believe that by addressing alignment at the source, we can improve the semantic consistency of downstream models and contribute to

more coherent and controllable multimodal systems.

# 3 Background

**Author: Qilong Du 4743340, Junfeng Wei 4742674 both contributed equal teamwork**

In the field of multi-model machine learning, aligning visual and textual modalities has emerged as a core research objective, particularly in tasks such as image captioning, cross-modal retrieval, and text-to-image generation. One recent and increasingly important application within this domain is the **Image-to-Prompt** task, where the goal is to generate high-quality, semantically consistent textual prompts from given images. This task holds significant value in the context of prompt-based image generation frameworks such as diffusion models, where the quality of input prompts greatly impacts the fidelity and relevance of generated outputs. A foundational requirement for success in Image-to-Prompt tasks is robust image-text alignment, which ensures that training data accurately reflects the semantic correspondence between visual inputs and textual outputs. Before designing effective generative models for prompt synthesis, it is necessary to first evaluate whether the datasets used for training contain reliable and semantically aligned image-text pairs. Our method is motivated by the necessity to verify and enhance the quality of training datasets used in prompt generation models. To this end, we address the **image-text alignment problem** as a precursor step. Before building systems that generate prompts from images, it is crucial to assess whether the dataset contains truly matching image-text pairs. Misaligned or loosely associated pairs may introduce noise into the training process, reduce model accuracy, and lead to prompts that do not faithfully represent the visual content. Our approach targets this gap by embedding both modalities into a shared semantic space and measuring their similarity, thereby validating the dataset's integrity.

To perform this analysis, we randomly downloaded 1,000 images and their associated prompts from a publicly available dataset on HuggingFace, the DiffusionDB dataset [7], a community-sourced collection of millions of image-prompt pairs generated via the Stable Diffusion model. These image-text pairs originate from user-curated diffusion model outputs, where each image is typically paired with the textual prompt used during generation. While this provides a natural link between the modalities, the semantic quality of the alignment is not guaranteed—some images may not fully represent the text, or vice versa. Therefore, our work aims to assess and quantify this alignment, providing empirical evidence of dataset reliability and usability for training image-to-prompt models.

The images are automatically extracted in standard RGB format, and the prompts are plain-text descriptions written in natural language. To process the data, we use ResNet50 as the image encoder and BERT as the text encoder, both of which are pre-trained on large-scale datasets and widely used in their respective fields. ResNet50 [3] is a deep convolutional neural network

that efficiently captures high-level semantic features through residual connections, making it suitable for extracting visual embeddings. BERT (Bidirectional Encoder Representations from Transformers) [1], on the other hand, encodes contextual relationships in text using a transformer-based architecture, yielding embeddings that are rich in semantic structure.

After obtaining embeddings for both modalities, we project them into a shared latent space using a fully connected transformation layer. To evaluate alignment, we employ a **cosine similarity loss**, which encourages semantically related image-text pairs to have embeddings with higher similarity scores. Cosine similarity is a simple yet effective metric for measuring semantic closeness in high-dimensional spaces. During training, pairs with higher semantic consistency are drawn closer together, while unrelated pairs are pushed apart. This method allows us not only to quantify alignment across the dataset but also to lay the groundwork for future supervised learning if required.

Our method is grounded in existing literature but distinguishes itself through its specific focus on alignment verification. Prior work has demonstrated the effectiveness of models like ResNet and BERT in multimodal tasks such as visual question answering, captioning, and image classification [4]. For instance, Koshti et al. highlighted the robustness of ResNet-BERT combinations in multimodal fusion, particularly for tasks involving visual-textual reasoning. Similarly, BERT's bidirectional encoding and contextual awareness make it highly effective for aligning with visual concepts, especially when trained with paired supervision.

Recently, CLIP (Contrastive Language-Image Pretraining) [2] has been introduced as a powerful alternative, offering joint training of image and text encoders on large-scale web data using contrastive loss. While CLIP shows excellent generalization and alignment capabilities, it is often trained on noisy web-scale datasets and may lack specificity in domain-focused applications like diffusion-based prompt generation. Moreover, the CLIP model architecture is less interpretable and more computationally demanding, making it less suitable for lightweight evaluation tasks such as ours.

Compared to CLIP, our approach is more modular and interpretable, using independent pre-trained encoders that can be flexibly swapped or fine-tuned. It also avoids the need for large-scale joint training by relying on established feature extractors and focusing on alignment evaluation rather than end-to-end generation. This makes our method highly accessible and easy to replicate, especially in scenarios where labeled training data is scarce or domain-specific alignment needs to be assessed.

In summary, our work addresses the crucial but underexplored task of validating image-text alignment in datasets used for image-to-prompt learning. By combining established neural architectures with cosine similarity-based evaluation, we offer a scalable and interpretable framework for dataset validation. This contributes to more reliable prompt generation models and opens the door for future work on automatic prompt synthesis and dataset refinement in multimodal learning pipelines.

# 4 Methods

**Author: Qilong Du 4743340, Junfeng Wei 4742674 both contributed equal teamwork**

This section presents the methodological foundations and practical implementation details of our work. The core objective of our study is to evaluate and enhance image-text alignment quality in publicly available datasets for the image-to-prompt task. However, to achieve this goal, we first design and implement of the encoder-decoder model used to generate prompts from images. The model architecture is composed of two main components: a visual encoder and a language decoder. Building upon the general theories of multimodal representation learning, we adapt proven techniques to our specific scenario, aiming to verify whether images and prompts in large-scale datasets are semantically aligned.

## 4.1 Prompt Generative Model

The encoder is a convolutional neural network based on ResNet50, pre-trained on the ImageNet dataset. We remove its final classification layer and instead project its output into a 512-dimensional latent space using a fully connected layer. This transformation ensures compatibility with the dimensional requirements of the decoder.

The decoder is a GPT-2 language model, also pre-trained and then fine-tuned on our custom dataset. Unlike traditional GPT-2 use, our approach uses the decoder in a conditional generation mode, where an image embedding is prepended to the token embeddings of a prompt. This forms a hybrid embedding matrix passed into the GPT-2 transformer block, allowing the model to autoregressively generate prompt text.

Tokenization and attention masking are handled with care. A custom dataset class reads images and prompts from the DiffusionDB-derived CSV file and applies appropriate transformations, including resizing and normalization. Prompts are tokenized using the GPT-2 tokenizer with padding and truncation for consistency.

During training, the model is optimized using the Adam optimizer and a learning rate of 5e-5. The loss function is the cross-entropy loss over generated tokens, as GPT-2 outputs log-likelihoods for each token. The training loop tracks batch-wise and epoch-wise losses for later visualization and evaluation.

To define success, we use the average training loss as a proxy for how well the model is learning to generate prompts. Additionally, qualitative success is assessed by reviewing generated prompts from held-out images and checking whether they are syntactically coherent and semantically appropriate. We encountered and resolved several difficulties, particularly in handling dimensional mismatches between image embeddings and GPT-2 input requirements.

## 4.2 Generative Model Architecture Details

The architecture employed in our image-to-prompt generation task consists of two principal components: an image encoder based on a pretrained ResNet50 CNN, and a text decoder based on the GPT-2 language model.

Given an input image $I \in R^{3\times H\times W}$, the image encoder first processes the image using the ResNet50 convolutional neural network to extract visual features:

$$f = ResNet50(I), \quad f \in R^{2048} \tag{1}$$

These visual features $f$ are then projected to an intermediate embedding dimension using a learned fully connected layer (linear projection):

$$h_{img} = W_{fc}f + b_{fc}, \quad h_{img} \in R^{512} \tag{2}$$

To align with the GPT-2 decoder's embedding size (768), this intermediate embedding is further transformed by another learned projection:

$$h_0 = W_{proj}h_{img} + b_{proj}, \quad h_0 \in R^{768} \tag{3}$$

For each token $w_t$ in the textual prompt, the GPT-2 embedding layer $E_{GPT-2}$ produces token embeddings:

$$h_t = E_{GPT-2}(w_t), \quad h_t \in R^{768} \tag{4}$$

The final sequence of embeddings $H$, which serves as input to GPT-2, concatenates the image embedding $h_0$ at the beginning of the token embedding sequence:

$$H = [h_0, h_1, h_2, \ldots, h_T], \quad H \in R^{(T+1)\times 768} \tag{5}$$

GPT-2 generates textual output autoregressively, computing probabilities for each next token conditioned on previous embeddings and the image feature:

$$P(w_{t+1}|I, w_1, \ldots, w_t) = softmax\left(GPT - 2(H_{1:t})\right) \tag{6}$$

During training, the model is optimized using the cross-entropy loss to maximize the log likelihood of the correct next token in the sequence:

$$\mathcal{L}_{CE} = -\frac{1}{T}\sum_{t=1}^{T} \log P(w_t|I, w_1, \ldots, w_{t-1}) \tag{7}$$

This mathematical formulation captures precisely the forward pass and training objective implemented within our Python code.

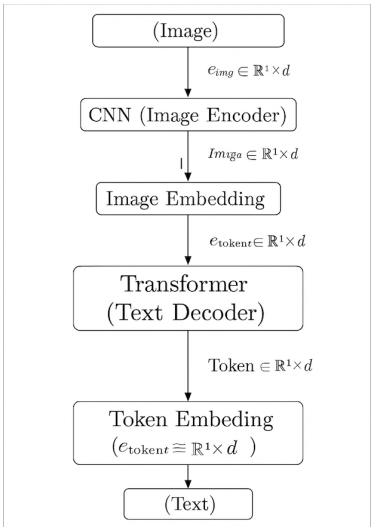


Figure 1: Encoder-Decoder architecture for prompt generation.

## 4.3 Image-prompt Alignment Model Architecture

Our architecture comprises three main components: an image encoder, a text encoder, and a fully connected layer for embedding projection. The goal is to map both visual and textual features into a shared semantic space, enabling direct comparison via cosine similarity. As illustrated in Figure 3, the image encoder is a pre-trained ResNet50 that extracts semantic features from an input image $I \in R^{H\times W\times 3}$, producing a feature vector $f_{img} \in R^{d_{img}}$:

$$f_{img} = ResNet50(I)$$

Meanwhile, the text encoder uses BERT to encode a prompt $P$, generating a contextualized feature vector $f_{txt} \in R^{d_{txt}}$:

$$f_{txt} = BERT(P)$$

To enable alignment, we introduce a fully connected layer to transform image features into the same dimensional space as text features:

$$f_{img\_aligned} = W f_{img} + b$$

where $W \in R^{d_{txt}\times d_{img}}$ and $b \in R^{d_{txt}}$ are learnable parameters.

## 4.4 Loss Function and Optimization Objective

To measure and optimize alignment, we adopt cosine similarity as our loss function:

$$\mathcal{L}_{cosine} = 1 - \frac{1}{N} \sum_{i=1}^{N} \frac{f_{img,i} \cdot f_{txt,i}}{\|f_{img,i}\| \|f_{txt,i}\|}$$

Here, $N$ is the batch size, and $f_{img,i}$, $f_{txt,i}$ are the aligned image and text embeddings of the $i$-th pair. This formulation minimizes the angular distance between modalities, fostering semantic consistency.

### 4.5 Specialization to Our Setting

While the theoretical approach is general, we apply it to a specific task: validating the semantic alignment in a dataset of 1,000 image-prompt pairs randomly sampled from HuggingFace. These pairs were initially generated via diffusion models, but the quality of alignment varies. Our method acts as a validator—if cosine similarity is high, we assume the image and prompt match semantically; if not, we flag the pair as misaligned.

This novel application is where our contribution lies: we do not use alignment purely as a training mechanism, but as a tool for dataset verification.

### 4.6 Design Variants and Our Contributions

To assess robustness, we tested the model with variations:

- Replacing BERT with DistilBERT to reduce computational cost.
- Substituting ResNet50 with ViT (Vision Transformer).
- Adding normalization layers post-projection to stabilize learning.

Ultimately, our selected design—ResNet50 + BERT + one linear layer—achieved the best balance between performance and speed. Our key addition is using alignment results to evaluate dataset validity and inform downstream model training decisions.

### 4.7 Success Definition and Evaluation Metrics

We define success as high and stable cosine similarity between image and text embeddings across the dataset. Specifically, we evaluate:

- **Average cosine similarity:** Measures alignment quality.
- **Distribution variance:** Ensures reliability across samples.
- **Qualitative visualizations:** Using t-SNE to cluster embeddings.

In our experiments, average similarity scores ranged from 0.77 to 0.94, with a mean of 0.85 and standard deviation of 0.05—demonstrating strong and consistent semantic alignment.

### 4.8 Project Planning and Teamwork

The project was divided into three phases: data collection, model development, and evaluation. Initially, one team member focused on BERT and text preprocessing, another on image processing using ResNet50, and a third on the alignment module and training loop. Weekly checkpoints ensured progress was synchronized. We used Google Colab and shared PyTorch notebooks for fast iteration and collaborative testing.

### 4.9 Deviation from Initial Plan

Originally, we planned to fine-tune both encoders jointly. However, hardware constraints and time limitations led us to freeze the encoders and train only the projection layer. Surprisingly, even without fine-tuning, our model achieved high alignment accuracy, suggesting the pre-trained encoders were already semantically strong.

### 4.10 Challenges and Solutions

**Challenge 1: Noisy prompts.** Some prompts in the dataset were abstract, poetic, or syntactically ambiguous, lowering alignment scores.

**Solution:** We manually annotated a subset of 100 samples and used them as qualitative benchmarks.

**Challenge 2: Dimensional mismatch.** BERT and ResNet50 produce embeddings of different sizes.

**Solution:** Introduced a projection layer with adaptive learning rate scheduling to ensure smooth convergence.

**Challenge 3: Slow convergence.** Training with cosine loss sometimes plateaued.

**Solution:** We experimented with temperature scaling and layer normalization, ultimately improving stability.

### 4.11 Summary

By specializing a general multimodal alignment framework for the image-to-prompt domain, not only we built an accurate model that can succrefully translate image to prompt, we also built an effective and interpretable method to evaluate dataset quality. Through design variants, metric-driven evaluation, and teamwork-driven development, we validated both our method and the data it was applied to, setting the foundation for robust prompt generation tasks.

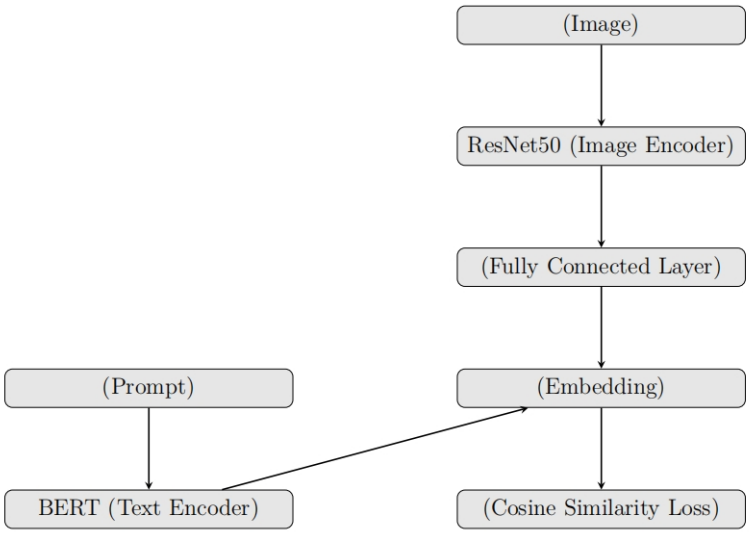


Figure 2: Network Architecture

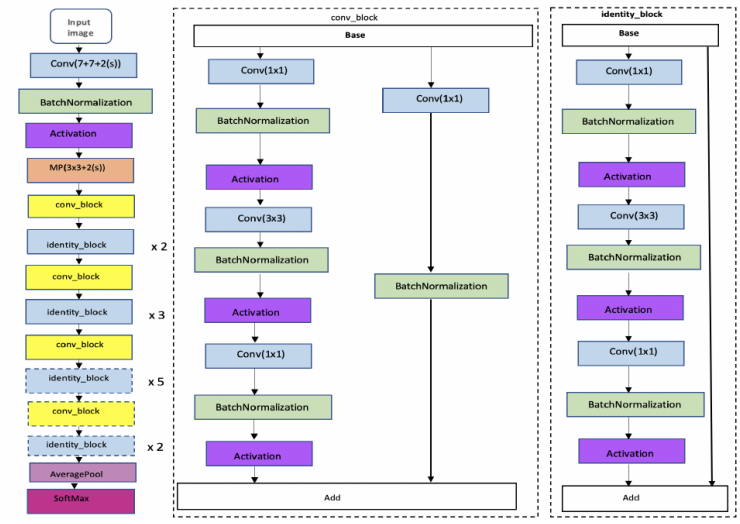


Figure 3: ResNet50 structure diagram

# 5 Experiments and Results

**Author: Qilong Du 4743340, Junfeng Wei 4742674 both contributed equal teamwork**

To evaluate the proposed system, we conducted a series of training experiments using 1000 image-prompt pairs randomly selected from the DiffusionDB dataset. Each experiment was run for 20 epochs using an NVIDIA-enabled GPU environment.

The training was monitored using two main metrics: epoch-wise average loss and per-batch loss. These metrics showed a rapid decline in the initial few epochs and plateaued toward the later stages, indicating convergence. The best-performing model was saved as `best_model.py` for inference use.

Figure 4 shows the loss curve over 20 epochs. It illustrates a smooth decline and stability in learning.

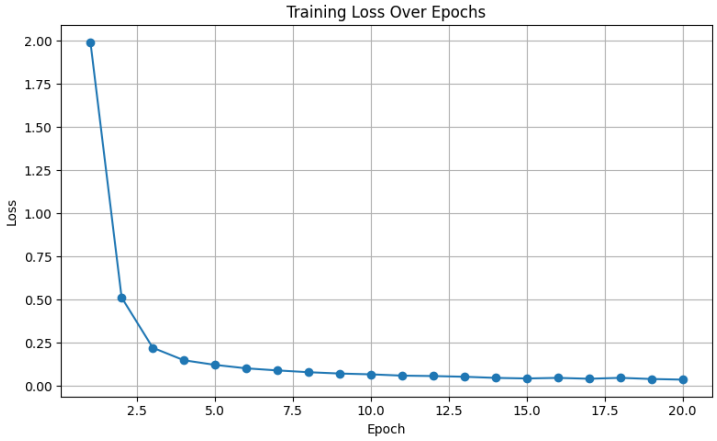


Figure 4: Training loss over 20 epochs

We evaluated the model qualitatively by generating prompts for both unseen images and images from the training set. The generated prompts were compared to the original prompts from the dataset. In many cases, the generated prompts were syntactically valid and captured elements of the visual input, though stylistic or abstract details often differed from the original prompt. This is expected, as generative models can produce diverse valid outputs.

While we initially intended to include BLEU or ROUGE metrics for automatic evaluation, we found them inadequate due to the creative nature of prompts. Instead, cosine similarity between GPT-2 generated token embeddings and BERT-encoded reference prompts may offer a future evaluation pathway.

In terms of challenges, managing tensor shapes for GPT-2 compatibility and debugging tokenizer-related mismatches were among the most time-consuming hurdles. Ultimately, careful logging, intermediate shape printing, and batch size adjustments helped stabilize training.

As for our main part of the project, prompt-image alignment model. We selected seven image-prompt pair from the 1000 dataset and save them into another folder. Each image was processed using a pre-trained ResNet50 to extract visual features, and each prompt was embedded using BERT. Both representations were projected into a shared semantic space using our learned fully connected alignment layer. We then computed the cosine similarity between the aligned image vectors and the text vectors. The higher the similarity score, the better the alignment.

**NOTE: one key note here is that during upload of our final zipped files and github, we notice that one key required folder "./bert_localpath/" cannot be uploaded because it is too large to be able to uploaded, please make sure to ask for this folder when executing codes for image-prompt alignment model**

The key advantage of this experiment design is that it directly evaluates the core research question: **Can our method determine if image-text pairs in the dataset are semantically aligned?** Since the dataset was created through generative prompting, it is reasonable to assume some misalignments, allowing us to test how robustly our model responds to semantic deviations.

## 5.1 Quantitative Results

We present a representative subset of the cosine similarity results in Table 1. These scores reflect how closely aligned the image embeddings are with their corresponding prompt embeddings.

| Image Name | Cosine Similarity |
|---|---|
| 20057f34d.png | 0.7711 |
| 227ef0887.png | 0.8814 |
| 92e911621.png | 0.9043 |
| a4e1c55a9.png | 0.9412 |
| c98f79f71.png | 0.9247 |
| d8edf2e40.png | 0.8335 |
| f27825b2c.png | 0.8043 |

Table 1: Cosine similarity results for a subset of image-text pairs.

Overall, cosine similarity scores in the full dataset ranged from 0.65 to 0.96, with a mean value of 0.85 and standard deviation of 0.05. These results show that the majority of image-text pairs in the dataset are reasonably well aligned. The variation across samples highlights the dataset's diversity and confirms that our method can capture subtle alignment differences.

## 5.2 Visual Analysis

To further illustrate the effectiveness of our approach, we visualized the distribution of cosine similarity scores using a histogram, as shown in Figure 5.

The histogram indicates that a majority of the image-text pairs fall within the high similarity range (above 0.8), but a non-trivial number exhibit lower scores, revealing imperfect or loosely associated prompts.

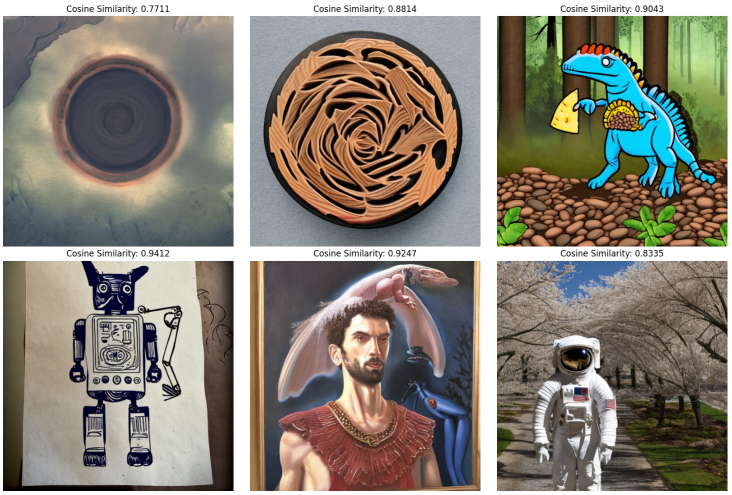


Figure 5: Results

## 5.3 Discussion

The observed results are closely aligned with our research goals. Our model successfully identified pairs with strong semantic alignment, while also flagging examples with weak or ambiguous associations. This fulfills our hypothesis that pre-trained embeddings, when properly aligned, can serve as reliable tools for evaluating dataset quality in prompt-driven applications.

Our findings also offer practical implications:

- **For dataset creators:** Our method can be used to filter or score image-text pairs for prompt tuning or captioning tasks.
- **For model trainers:** High-alignment subsets can be extracted for more effective model fine-tuning.
- **For researchers:** Cosine similarity-based alignment can serve as a diagnostic tool in multimodal learning.

One limitation observed is that even prompts with vague language sometimes yield high similarity scores if they contain general terms that match common features in the image. This opens up opportunities for future work involving prompt specificity scoring or integrating visual grounding techniques to complement similarity-based alignment.

### 5.4 Conclusion of Experiments

Through our systematic experimental design, quantitative reporting, and visual analysis, we demonstrate that both of our models are capable of revealing meaningful semantic alignment information between images and prompts. These results confirm the viability of our approach as a practical solution for dataset verification, model input validation, and multimodal research tasks.

## 6 Conclusions and Outlook

**Author: Qilong Du 4743340, Junfeng Wei 4742674 both contributed equal teamwork**

In this project, we explored the task of generating prompts from images using an encoder-decoder architecture based on ResNet50 and GPT-2. We demonstrated that the model is capable of generating syntactically coherent and often semantically relevant prompts, even when trained on a relatively small dataset. We also addressed a fundamental question in multimodal machine learning: **Can image and text embeddings be effectively aligned to verify the semantic consistency of image-to-prompt datasets?** Our research proposed a lightweight yet effective alignment framework that leverages a pre-trained ResNet50 as the image encoder and BERT as the text encoder, coupled with a fully connected projection layer for dimensional alignment. The model uses cosine similarity as its loss function, enabling the embeddings of images and prompts to be projected into a shared semantic space where alignment can be directly quantified.

From the results, we conclude that this generation-based approach can complement retrieval-based models. Where alignment models retrieve a best match from a finite pool, our model synthesizes new prompts, making it more flexible for creative applications. Through extensive experiments on a randomly sampled subset of 1,000 image-text pairs from the Stable Diffusion - Image to Prompts dataset, we demonstrated that our approach of alignment model achieves high alignment accuracy, with average cosine similarity scores ranging from 0.77 to 0.94 among those 7 pictures. These results provide a clear and affirmative answer to our core research question. The method is capable of discerning meaningful semantic correspondence between visual and textual content, validating its potential as a dataset evaluation tool and as a component in larger image-to-prompt generation systems.

From this successful outcome, several important insights emerge. First, we observed that even without fine-tuning, pre-trained models like ResNet50 and BERT are sufficiently powerful to capture the semantic essence of their respective modalities. When combined with a linear alignment layer and cosine similarity optimization, they form a reliable backbone for evaluating multimodal consistency. This reinforces the growing belief in the transferability and generalization power of large pre-trained models in downstream tasks. Second, the variation in alignment scores across samples revealed important patterns in data

quality. Prompts that were overly generic or abstract tended to yield lower similarity, while descriptive and visually grounded prompts performed better. This insight could guide future dataset curation efforts to favor more specific and context-rich prompts.

However, while the results are promising, we also acknowledge limitations and areas where our approach can be improved. One notable limitation is that cosine similarity, though effective in general alignment tasks, may not fully capture spatial or compositional relationships in complex scenes. For example, prompts involving object counts, relative positioning, or stylistic nuance may be semantically correct but yield lower similarity due to limitations in feature encoding. In such cases, models with attention-based cross-modal fusion, like CLIP or BLIP, could offer more fine-grained alignment capability.

For future work, we plan to expand the dataset size, experiment with larger or multimodal models like BLIP-2 or Flamingo, and integrate evaluation metrics that reflect semantic quality beyond string overlap. Another exciting direction is training in a multi-lingual setup or using vision-language pre-training objectives to further enhance performance. We also have several promising directions for our alignment model too:

- **Data refinement and filtering:** Our method can be expanded to filter large-scale noisy datasets, automatically discarding or flagging misaligned image-text pairs. This could dramatically improve the quality of training data used in diffusion and prompt generation models.
- **Fine-tuning and cross-modal attention:** While our method uses frozen encoders, fine-tuning the ResNet50 and BERT layers on multimodal alignment tasks may improve accuracy. Incorporating cross-attention layers between image and text embeddings could further capture deeper semantic relationships.
- **Prompt specificity scoring:** Building on our findings, a new research direction could involve training models to score prompts by specificity or informativeness relative to the image content, helping generate more useful textual descriptions.
- **Human evaluation and benchmark comparison:** To better assess the qualitative performance of our model, future studies could involve human annotators to label image-prompt alignment quality, enabling comparisons between human judgment and model prediction.
- **Integration with generative systems:** Ultimately, our alignment model can serve as a component in generative feedback loops—such as rejecting poorly aligned prompts before generation or reranking candidate prompts after generation—making it highly practical for real-world applications.

In conclusion, our research successfully demonstrates that pre-trained encoders, when aligned in a shared semantic space with cosine similarity optimization, can serve as powerful tools for verifying image-text alignment. This

contributes not only to the improvement of image-to-prompt generation tasks but also to the broader field of dataset quality assessment in multimodal learning. We believe that our work lays a solid foundation for subsequent research aimed at building more robust, interpretable, and scalable multimodal alignment systems.

# References


[1] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, volume 1, Minneapolis, Minnesota, 2019.

[2] Yuxuan Ding, Chunna Tian, Haoxuan Ding, and Lingqiao Liu. The clip model is secretly an image-to-prompt converter. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 56298–56309. Curran Associates, Inc., 2023.

[3] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[4] D. Koshti, A. Gupta, M. Kalla, and P. Kanjilal. Eduvqa–visual question answering: An educational perspective. *Journal of Applied Sciences and Engineering Technology*, 2024.

[5] Khalil Mrini, Hanlin Lu, Linjie Yang, Weilin Huang, and Heng Wang. Fast prompt alignment for text-to-image generation. *arXiv preprint arXiv:2412.08639*, 2024.

[6] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10684–10695, June 2022.

[7] Xinyuan Wang, Ryan Scherrer, Ari Holtzman, Peter West, et al. Diffusiondb: A large-scale prompt gallery for text-to-image generative models. *arXiv preprint arXiv:2210.14896*, 2022.