Hybrid Approach to Vision–Language Generation: Custom Architectures and Fine-Tuning

BADRI Insaf , SGHIR Marwa

Cherradi Mohammed

National School of Applied Sciences of AL Hoceima
Abdelmalek Essaâdi University, MOROCCO

Thursday 22, June 2025

ABSTRACT

The interplay between vision and language has led to major advances in both image-to-text and text-to-image generation. In this research, we explore both directions by addressing the challenges of image captioning and personalized image generation. First, we present a comprehensive survey of image captioning, the task of generating natural language descriptions for visual content, which has become a key area in computer vision and NLP. Deep learning models, particularly convolutional neural networks (CNNs) and recurrent neural networks (RNNs), have significantly advanced the field by improving the accuracy and contextual relevance of generated captions. We provide a systematic analysis of deep-learning-based approaches, focusing on prominent architectures such as ResNet-50 and DenseNet-201. Commonly used datasets, including MS COCO and Flickr8k, are reviewed alongside standard evaluation metrics. In addition, we examine recent developments in vision-language pre-training models and their impact on captioning performance. To demonstrate these improvements, we fine-tuned the BLIP model on the COCO dataset, achieving notable gains in caption quality. In the second part of this work, we address the challenge of personalizing large-scale diffusion models like Stable Diffusion XL. We propose an efficient fine-tuning strategy that combines DreamBooth with Low-Rank Adaptation (LoRA), enabling adaptation of the model to a curated dataset of 3D icon images. The fine-tuned model effectively preserves stylistic fidelity while generalizing semantic content, offering a scalable and reproducible solution for personalized image generation in low-data settings. Together, these contributions highlight promising directions for advancing both descriptive and generative capabilities in vision-language systems.

Keywords Image Captioning, Deep Learning, Neural Networks, NLP, Resnet50 densenet201, VisionLanguage , Image generation ,vision-language models, convolutional neural networks , recurrent neural networks , BLIP model ,Stable Diffusion XL DreamBooth Low Rank Adaptation , diffusion models.

I. Introduction :

Automatic image captioning is a challenging research area with broad applications, ranging from human-computer interaction and medical imaging to industrial inspection and assistive tools for the visually impaired. By combining computer vision and natural language processing, it generates meaningful text descriptions for images without human input. The goal is to take an input image and generate a caption that describes its content. Recent deep learning advances have significantly improved model performance, enabling wider adoption across real-world systems. Similarly, recent advancements of text-to-image generation capabilities, including several diffusion models like Stable Diffusion XL [1], have ushered in a new era for AI-assisted creation. However, the high cost and demands from these models make them difficult to fine-tune for a niche application. This work seeks to enhance the finetuning of models to create domain-specific images - addressing the high costs and demands from current text-to-image diffusion models.

In image captioning, researchers are usually confronted with a set of problems, some of which are commonly experienced in many artificial intelligence tasks. These include the exposure bias problem, the loss-evaluation mismatch problem, the vanishing gradient problem, and the exploding gradient problem. In this project, we encountered several challenges, including addressing the model's ability to effectively capture the semantic relationships between consecutive words in the generated captions. While diffusion-based models like Stable Diffusion XL produce impressive text-to-image generation results, it is not always feasible to adapt them to a specific domain (i.e., when large datasets and the compute resources may not be available). Sometimes there are simply no options to customize them, which limits the utility of a model to use in lightweight or specialized contexts and where labeled training data may be scarce. Semi- and fully supervised personalization approaches still have drawbacks like overfitting, broken semantic relationships, and slow inferences that are often left unresolved in many real-world cases.

To address these problems, we present a novel approach for combating the challenges of image captioning by integrating Convolutional Neural Networks (CNNs) employing two methodologies: ResNet50 and DenseNet210. These two approaches, each joined with their own Long Short-Term Memory networks (LSTMs), aim to strengthen the performance of the given model. The BLIP approach has also been fine-tuned on a subset of the COCO dataset to enhance the precision for generating accurate and contextually meaningful image descriptions, thereby improving the overall quality and coherence of the generated descriptions. In an attempt to effectively customize large-scale diffusion models, we provide a lightweight fine-tuning method that integrates DreamBooth and Low-Rank Adaptation (LoRA). This method enables effective adaptation for the Stable Diffusion XL model using only a few images, significantly reducing the computational and memory requirements compared to traditional fine-tuning. Our method stands out in two key aspects. First, it facilitates unrestricted personalization on limited data while preserving semantic fidelity and generative quality. Second, it employs LoRA's efficient parameter reduction to sustain low-cost training on small computing devices, thus meeting real-world or resource-limited conditions. This approach is tested on a curated 3D icon dataset, showcasing the model's ability to integrate style transfer while maintaining novel visual generation and surpassing the training exemplars.

II. Previous Related Work

Image captioning has evolved significant attention in recent years of deep learning, particularly through the integration of computer vision and natural language processing. Early models relied on handcrafted features or template-based approaches, but the field shifted dramatically with the introduction of **encoder-decoder** architectures. In these models, (CNNs) like **VGG16** and **ResNet** were used to extract visual features, while (RNNs), especially (LSTM) units, generated textual descriptions. The *Show and Tell* model (**Vinyals et al., 2015)** was among the first to demonstrate end-to-end learning for caption generation. This was later extended by the *Show, Attend and Tell* model **(Xu et al., 2015),** which introduced visual attention mechanisms to dynamically focus on different regions of an image during caption generation.

Recent advances have leveraged Transformer-based models and multimodal pretraining. **VisionEncoderDecoder** architectures, such as **ViT-GPT2**, combine a Vision Transformer (ViT) as the encoder with a pretrained language model like GPT-2 as the decoder, enabling more coherent and contextually rich captions. Similarly, models like BLIP and OFA utilize large-scale vision-language pretraining and demonstrate strong performance on benchmark datasets. Fine-tuning these models on datasets such as **COCO 2017**. These advancements underscore the importance of attention mechanisms, transformer-based architectures, and large-scale pretraining in pushing the state-of-the-art in image captioning.

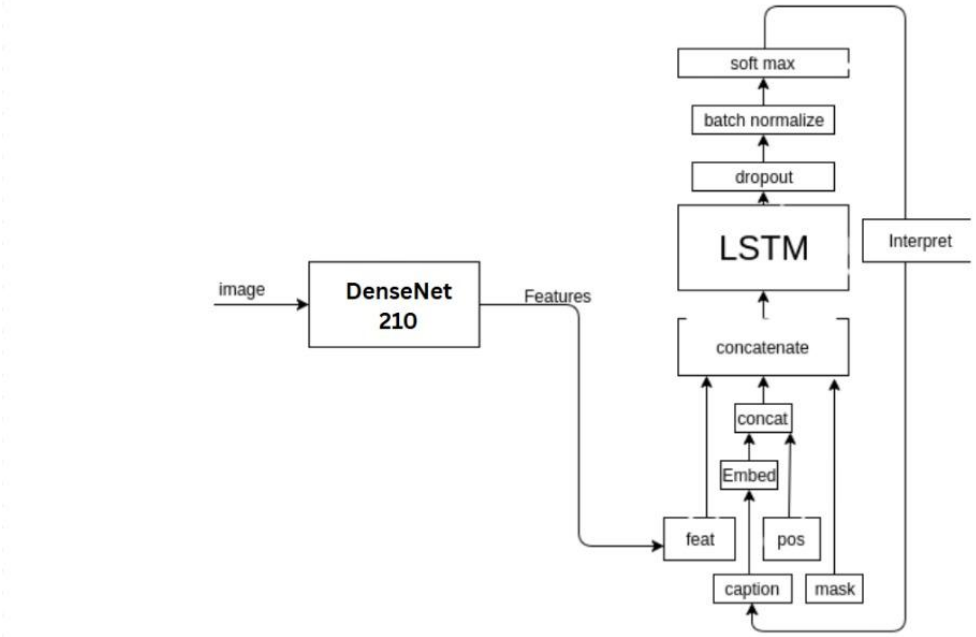
Foundational Research Papers :

Paper	**Contribution**	**Year**
Show and Tell	Introduced CNN-RNN encoder-decoder architecture for end-to-end image captioning	**2015**
Show, Attend and Tell	Added attention mechanism to focus on image regions while generating captions.	**2015**
Bottom-Up and Top-Down Attention	Improved attention using object detection (Faster R-CNN).	**2018**
OSCAR	Vision-language pretraining using object tags (visual grounding).	**2020**
BLIP	Unified vision-language model pretraining for image-text tasks.	**2022**
ViT-GPT2	Vision Transformer + GPT-2 decoder used in encoder-decoder architecture	**2021 -present**

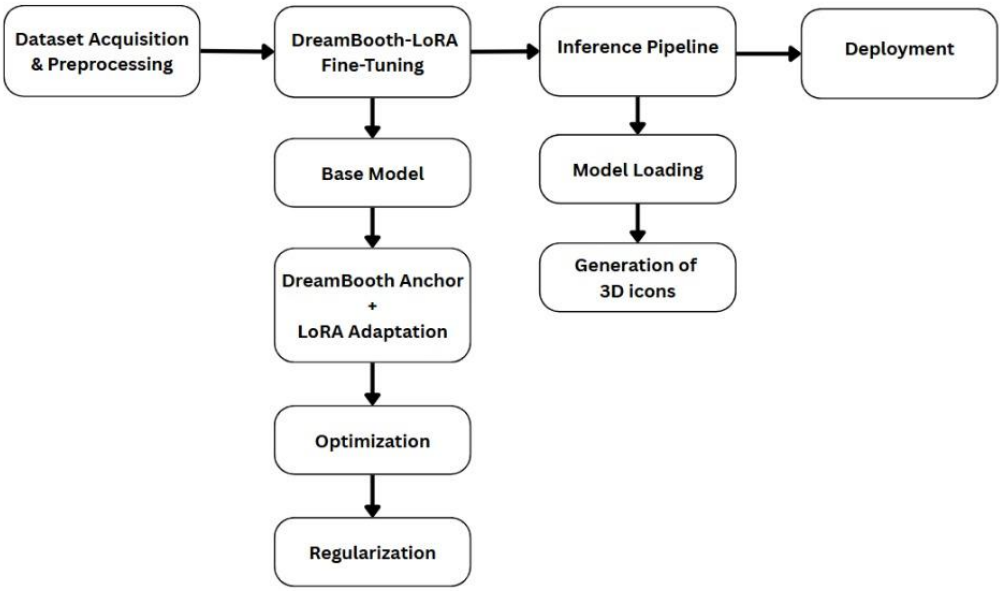
The concurrent work of **Gal et al.** [2] proposes a method to represent visual concepts, like an object or a style, through new tokens in the embedding space of a frozen text-to-image model, resulting in small personalized token embeddings. While this method is limited by the expressiveness of the frozen diffusion model, our fine-tuning approach enables us to embed the subject within the model's output domain, resulting in the generation of novel images of the subject which preserve its key visual features.

Diffusion Probabilistic Models (DPMs) [3] have recently achieved state-of-the-art results in density estimation and image generation quality, especially when using **UNet** backbones and reweighted training objectives. Despite their strengths, DPMs face high training costs and slow inference due to operating in pixel space. Latent Diffusion Models (LDMs) address these issues by working in a lower-dimensional latent space, significantly reducing computational cost with minimal loss in quality. This efficiency is central to the **Stable Diffusion XL** model used in our fine-tuning pipeline.

III. Methodology



DenseNet 210 LSTM architecture for image captionating



DreamBooth-LoRA Fine-Tuning Pipeline for 3D Icon Style generation

Figure 1: Data Flow Diagram

- **IMAGE-TO-TEXT**

1. Dataset collection

There are many open-source datasets for tackling image captioning challenges available, and some of them are Flickr8k containing 8,000 images. In this case study, our selection has been Flickr8k dataset. It consists of 8,000 images each having five corresponding captions for test purposes. The captions in this dataset are annotated by hand and are both descriptive and informative in nature. From literature reviews, it has already been used extensively for image captioning studies. It has a prominent advantage as it allows researchers to train and test their models on a large variety of visual materials and present a huge and varied set of images. On the other hand, we employed a subset of COCO dataset for fine-tuning on BLIP model.

2. Techniques in Image Captioning

ResNets are employed in object detection and are inspired by the pyramidal cells in the brain's cerebral cortex, utilizing slip connections to interlink multiple layers. Typically, ResNets use two or three skip connections and are constructed by stacking "Residual Blocks"for instance, ResNet-50 comprises 50 such layers. This architecture facilitates faster and easier optimization compared to traditional networks lacking skip connections and residual blocks. Several studies in this survey have used ResNet for object detection and generating image representations. **DenseNets**, another type of CNN, leverage dense connections via Dense Blocks, linking all layers with similar feature-map sizes directly to one another. Each layer receives inputs from all preceding layers and forwards its output to all subsequent layers, maintaining a feed-forward flow. RNNs, which incorporate internal memory, process sequential data where each output depends on previous computations. Though suitable for tasks like handwriting and speech recognition, RNNs struggle with vanishing gradients and handling long sequences effectively. To address these challenges, (LSTMs) were introduced. Designed to capture long-term dependencies, LSTMs use gate mechanisms to regulate information flow and have been applied to complex problems such as handwriting generation, language modeling, speech synthesis, and video analysis. Despite their effectiveness, LSTMs require significant memory and overlook hierarchical sentence structure. They are widely adopted in Encoder-Decoder frameworks for image captioning, enabling the generation of textual representations from visual data.

3. Model architecture

3.1 CNN-LSTMS approch

The model has an encoder-decoder architecture. The encoder employs a pre-trained DenseNet-201 and other hand Resmet50 trained on ImageNet to extract visual features from the input images. The last classification layer of DenseNet is dropped to learn a fixed-size visual feature vector or spatial feature map for the image. The decoder has a language model based on an LSTM. It has an embedding layer used for converting the indices of words into dense vectors and an LSTM for processing the embedded words sequence and the image features together. It has a final fully connected layer mapping the output of an LSTM to the vocabulary space used for prediction of the next word in the caption sequence.

3.2 Fine-tuned approach

For this approach, we use the BLIP model, which combines a Vision Transformer to understand images and a Transformer-based decoder that writes captions by looking at both the image features and what it has already generated. We fine-tune the entire model on a smaller portion of the COCO dataset, teaching it to create accurate and meaningful captions. To help the model learn effectively without overfitting, we use techniques like adjusting the learning rate over time. Before feeding images into the model, we resize and normalize them so they match the expected input format, and we carefully tune training settings like batch size and learning rate to get the best result

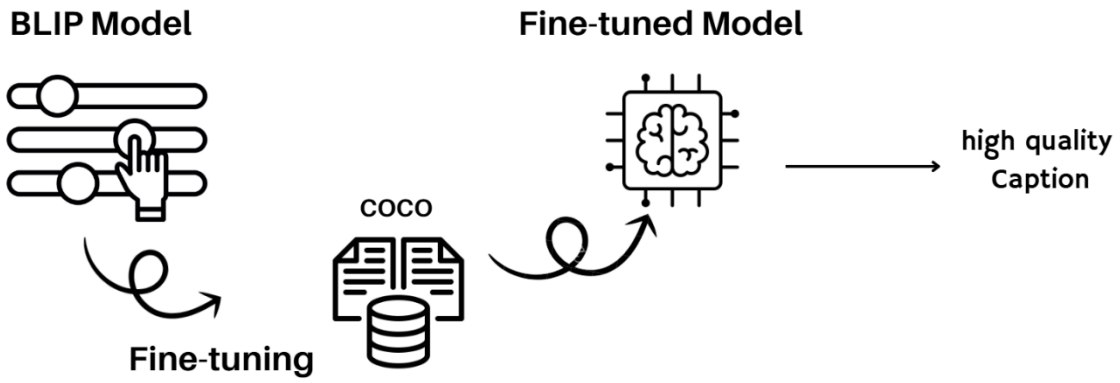


Figure 2: Fine-tuned model architecture

4. Evaluation Metrics for Image Captioning Methods:

Evaluation metrics in image captioning are generally divided into two categories: text evaluation metrics and caption evaluation metrics. Text evaluation metrics assess machine-generated text independently and were initially developed for evaluating outputs from machine translation systems. In contrast, caption evaluation metrics are specifically designed for evaluating the quality of image-generated captions. One widely used text evaluation metric is **BLEU** (Bilingual Evaluation Understudy), which compares segments of generated text against reference segments and calculates an average score based on n-gram overlaps, though it does not account for syntax. BLEU remains popular due to its language independence, simplicity, speed, low cost, and strong alignment with human evaluation. Another important metric is ROUGE (Recall-Oriented Understudy for Gisting Evaluation), which is often used in text summarization tasks. ROUGE measures the overlap of n-grams, word sequences, and word pairs between machine-generated summaries and human-written references. Among its variants, ROUGE-N specifically quantifies n-gram overlap, offering a clear measure of the informativeness and relevance of generated text.

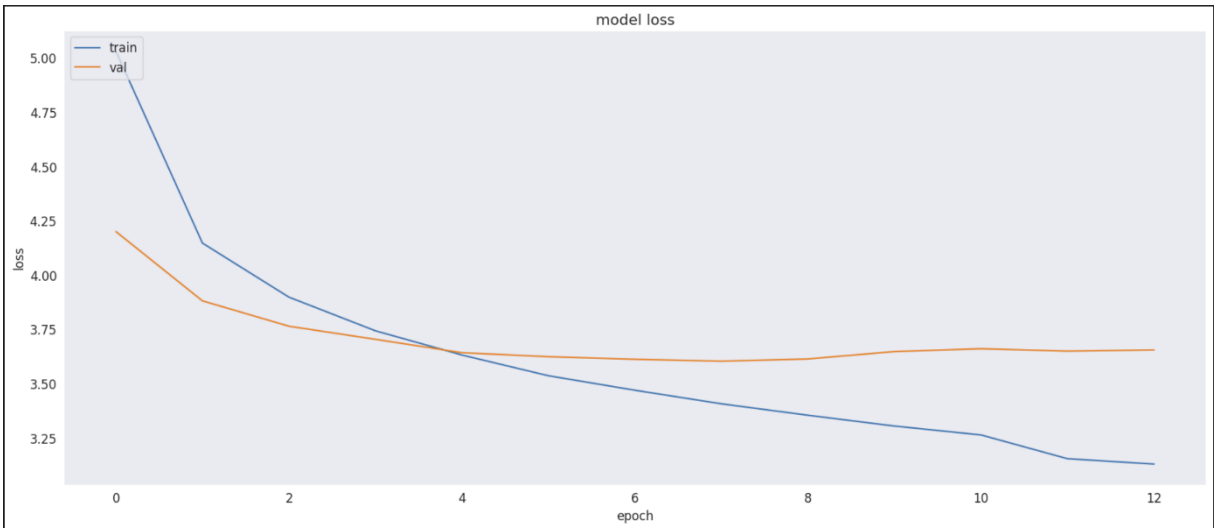


Figure 3: Learning curve of the model

- **TEXT-TO-IMAGE**

This project follows a structured pipeline designed to efficiently fine-tune a large-scale text-to-image diffusion model on a small, domain-specific dataset. Our workflow consists of four main stages: dataset preparation and prompt generation, base model setup, fine-tuning with DreamBooth and Low-Rank Adaptation (LoRA), and inference and visual evaluation. The objective is to adapt the Stable Diffusion XL (SDXL) model to generate novel 3D icons that maintain the visual style and semantic structure of the training data, without requiring large computational resources.

1. Stable Diffusion Models: Background

Stable Diffusion is a class of text-to-image generative models based on **latent diffusion** techniques. Instead of generating images directly in pixel space, these models operate in a **compressed latent space** learned by a Variational Autoencoder (VAE). A text encoder (e.g., CLIP) converts natural language prompts into embeddings that guide the image synthesis process through a **denoising UNet architecture**.

The model starts from random Gaussian noise and iteratively denoises it over multiple steps to produce a coherent image that aligns with the input text. This design leads to significant computational efficiency and high-resolution image synthesis compared to earlier diffusion models like DALL·E 2 or Imagen.

In particular, **Stable Diffusion XL (SDXL)**, the latest version of this model family, improves generation quality through:

- A dual text encoder setup (CLIP and T5),
- A more expressive latent space,
- Improved prompt understanding and image coherence,
- High-resolution (1024×1024) outputs.

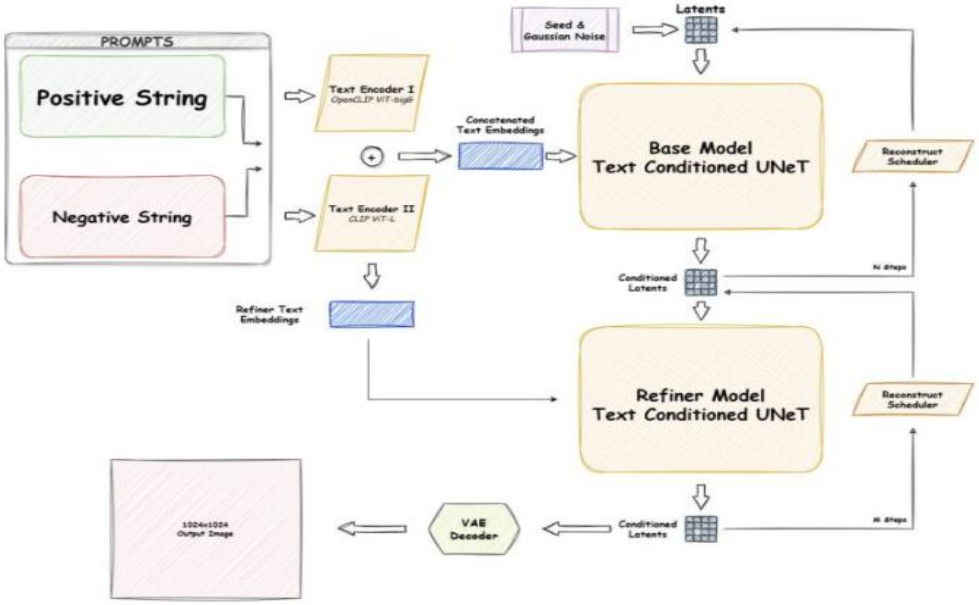


Figure 4: Architecture of stable Diffusion XL

2. Dataset Preparation and Captioning

We used a curated dataset of 3D icon images available on the Hugging Face Hub under the repository ID linoyts/3d_icon [4]. The dataset consists of high-resolution 3D icons. Since the dataset did not include textual captions, we employed the BLIP (Bootstrapping Language-Image Pretraining) model to automatically generate descriptive captions for each image. This step was essential to enable supervised fine-tuning of the diffusion model via DreamBooth.

Each image-caption pair was formatted and resized to 1024×1024 pixels and stored locally in a directory compatible with the Hugging Face training pipeline. The captions served as prompts during fine-tuning and inference.

Figure 5: Preview of the Dataset

3. Model and Training Configuration

Our work builds on stabilityai/stable-diffusion-xl-base-1.0, a cutting-edge latent diffusion model capable of generating high-resolution images from text. To improve the quality and consistency of decoded outputs, we integrated a more stable Variational Autoencoder (VAE) from madebyollin/sdxl-vae-fp16-fix.

To enable efficient fine-tuning on a small dataset, we combined two techniques: **Low-Rank Adaptation (LoRA)** and **DreamBooth**.

LoRA reduces the number of trainable parameters by introducing low-rank matrices into attention layers, making fine-tuning feasible on limited hardware.

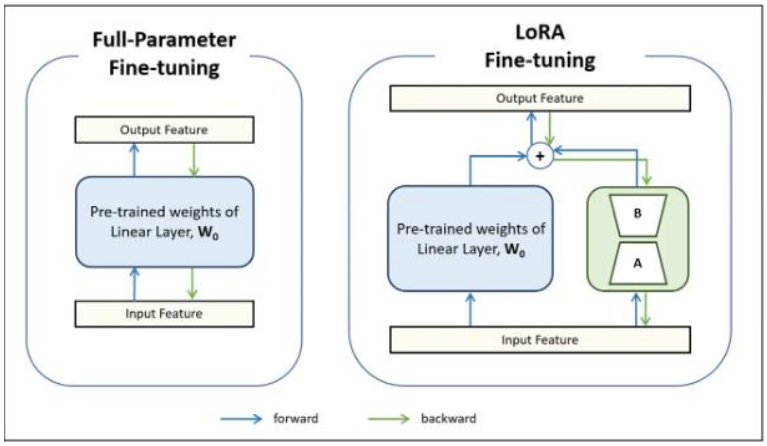


Figure 6: Lora Fine-tuning Concept

DreamBooth, on the other hand, enables the model to associate a visual concept with a unique trigger token—allowing the model to personalize generation outputs based on a small number of example images.

All model configuration and training were conducted using the diffusers, peft, and accelerate libraries from Hugging Face, which provided a modular and efficient training framework.

4 Fine-Tuning Procedure

Fine-tuning was carried out using the train_dreambooth_lora_sdxl.py script, which integrates both DreamBooth and LoRA in a unified training pipeline. We supplied the local dataset directory, defined "prompt" as the caption column, and configured LoRA-specific parameters such as rank, learning rate, and number of training steps.

```
--pretrained_model_name_or_path="stabilityai/stable-diffusion-xl-base-1.0" \
--pretrained_vae_model_name_or_path="madebyollin/sdxl-vae-fp16-fix" \
--dataset_name="./3d_icon_dataset/" \
--caption_column="prompt" \
--output_dir="3dicon_lora_model" \
--mixed_precision="fp16" \
--instance_prompt="a 3D icon in the style of TOK" \
--resolution=1024 \
--train_batch_size=1 \
--gradient_accumulation_steps=4 \
--gradient_checkpointing \
--learning_rate=1e-4 \
--snr_gamma=5.0 \
--lr_scheduler="constant" \
--lr_warmup_steps=0 \
--use_8bit_adam \
--max_train_steps=500 \
--checkpointing_steps=500 \
--seed="0"
```

Figure 6: Training Parameters for the Models

To teach the model the visual style of 3D icons, we used a consistent prompt format: "a 3D icon in the style of TOK", where TOK served as the trigger token. This allowed DreamBooth to capture and encode the unique aesthetic of our dataset. LoRA ensured that only a minimal subset of model parameters were fine-tuned, significantly reducing memory consumption and training time.

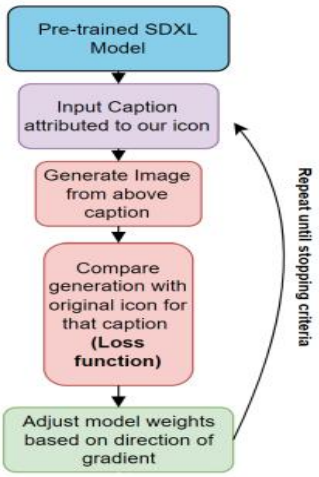


Figure 7: A visual representation of how the model learns to generate images according to a style

IV. Results

- IMAGE-TO-TEXT :

To evaluate the performance of our image captioning model, we conducted both quantitative analysis using standard NLP metrics and qualitative assessment through visual inspection of generated captions. The results demonstrate how well the model understands visual content and translates it into coherent natural language.

Quantitative Results

Metrics	**Score**
BLEU-1	0.40
BLEU-2	0.35
ROUGE	0.30

Le score BLEU-1 de 0.40 indique que 40 % des unigrammes générés par le modèle correspondent aux mots des légendes de référence, ce qui montre une capacité raisonnable du modèle à produire des mots pertinents. Le score BLEU-2 de 0.35, légèrement élevé, révèle que le modèle est également capable de produire des paires de mots bigrams cohérentes avec les légendes attendues, traduisant une certaine fluidité syntaxique dans les phrases générées. En revanche, le score ROUGE-1 de 0.30 suggère que seulement 30 % des mots présents dans les légendes de référence sont retrouvés dans les légendes générées, ce qui montre une couverture lexicale plus limitée. Globalement, ces résultats indiquent que le modèle capture assez bien le contenu visuel et est capable de le transformer en phrases compréhensibles.

- TEXT-TO-IMAGE :

To evaluate the effectiveness of fine-tuning the base Stable Diffusion XL (SDXL) model with DreamBooth and LoRA on a 3D icon dataset, we performed a series of qualitative and

quantitative analyses. These aimed to assess semantic alignment and style transfer fidelity, using neutral prompts and various perceptual similarity metrics.

→ **Qualitative Results**

We compared outputs from the base SDXL model and the fine-tuned model using identical prompts.

For instance the following prompts were given:

- **Base modal's Prompt**: "a 3D icon of a Marshmallow"
- **Fine-tuned model's Prompt**: "a 3D icon of a Marshmallow in **the style of TOK**"

Figure 8: Baseline generation

Figure 9: Fine-tuned generation

These comparisons visually demonstrate that the fine-tuned model has effectively incorporated stylistic features from the 3D icon dataset.

→ **CLIP-Based Semantic Evaluation**

To assess semantic alignment between generated images and text prompts, we used the CLIP model. Specifically:

- **CLIP Score**: Measures the alignment between a generated image and a guiding text prompt. A score close to 1.0 indicates strong semantic alignment.

We evaluated three prompt-image pairs and obtained the following scores:

- Prompt 1: **1.0**
- Prompt 2: **1.0**
- Prompt 3: **1.0**

These perfect scores indicate that the generated images are strongly aligned with their respective prompts.

→ **Style Similarity with Cosine Similarity**

In addition to semantic similarity, we evaluated style similarity by comparing the cosine similarity between the CLIP embeddings of generated and reference icons.

- **Cosine Similarity of CLIP Image Embeddings:** Measures the similarity between the embedding of a style reference image and its corresponding generated image. This helps evaluate how closely the generated image matches the style of the target dataset.

We evaluated three images and obtained the following scores:

- Example 1: 0.537
- Example 2: 0.572
- Example 3: 0.523

These moderate similarity values suggest that the model captures key stylistic features of the 3D icon dataset, though not perfectly.

→ **Gram Matrix Style Analysis**

We further evaluated style fidelity using Gram matrix comparisons (from VGG19 feature maps). The mean squared error MSE between the Gram matrices of generated and reference images were:

Case 1: 1.04×10^{-5}

Case 2: 1.31×10^{-5}

Case 3: 3.03×10^{-6}

Low Gram loss values indicate that the fine-tuned model closely captures the stylistic texture and layout of the reference images.

➢ **Interpretation of results:**

These results collectively demonstrate that the fine-tuned LoRA model outperforms the base model in adapting to the stylistic features of 3D icons, while preserving semantic fidelity to the prompts.

The slightly moderate cosine similarity suggests potential room for improvement in capturing fine-grained stylistic details. This may be due to the limited size of the training dataset or the expressiveness of the LoRA layers compared to full fine-tuning approaches.

Nevertheless, the consistency in CLIP scores and low Gram losses support the hypothesis that LoRA fine-tuning effectively enables personalized image generation in low-resource scenarios.

V. Conclusion

On one hand Image captioning has seen major breakthroughs in the past few years. Recent advances based on deep learning techniques have significantly improved the accuracy of generated captions. Textual descriptions of images enhance the efficiency of content-based image retrieval and broaden the applicability of visual understanding in fields such as medicine, security, and the military, offering strong potential for practical deployment. Additionally, the theoretical models and research methodologies developed for image captioning contribute to the advancement of related areas like image annotation, visual question answering (VQA), cross-media retrieval, video captioning, and video dialogue, highlighting its substantial academic and practical significance. On the other hand, we presented a fine-tuning pipeline using DreamBooth LoRA to adapt the base Stable Diffusion XL model to a custom dataset of 3D icons, with the main objective of enabling the model to generate high-quality 3D-style icons from text prompts using only a limited number of training examples. Our evaluation, based on both qualitative inspection and quantitative metrics, demonstrates that the fine-tuned model preserves the semantic content of prompts while successfully adopting the visual style of the 3D icon dataset. CLIP scores of 1.0 confirm perfect text-image alignment, and low Gram losses

indicate strong stylistic consistency with reference images. Cosine similarities between generated and original icons suggest a moderate match in visual features, highlighting potential for further stylistic refinement. These results support our hypothesis that LoRA fine-tuning can effectively personalize image generation using a lightweight, resource-efficient process, although improvements may be achieved with a more diverse or larger dataset.

VI. Future Work

Future work could involve testing the fine-tuned model on broader icon categories or alternative 3D styles to assess generalization. Additionally, integrating user-in-the-loop feedback or reinforcement learning could help improve stylistic accuracy over time. Finally, evaluating LoRA against other parameter-efficient fine-tuning methods (like full DreamBooth or Textual Inversion) could offer deeper insight into trade-offs between quality, compute, and data requirements.

The current CNN-LSTM image captioning model can be improved by integrating attention mechanisms to let the model focus on relevant image regions during caption generation, resulting in more accurate descriptions. Additionally, replacing the LSTM decoder with Transformer-based architectures such as ViT combined with Transformer decoders or vision-language pretrained models like CLIP and Flamingo can significantly boost caption fluency and relevance, reflecting recent state-of-the-art advancements.

VII. Acknowledgment

We would like to express our sincere gratitude to our supervisor, **Dr. Mohamed CHERRADI**, for his valuable guidance, insightful feedback, and continuous support throughout the course of this work. His expertise and encouragement were instrumental in the successful completion of this project

VIII. References

[1]: D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. M¨uller, J. Penna, and R. Rombach, “SDXL: Improving Latent Diffusion Models for High-Resolution Image Synthesis,” arXiv preprint arXiv:2307.01952, 2023. [Online]. Available: https://arxiv.org/abs/2307.01952.

[2]: Rinon Gal, Yuval Alaluf, Yuval Atzmon, Or Patashnik, Amit H Bermano, Gal Chechik, and Daniel CohenOr. An image is worth one word: Personalizing text-toimage generation using textual inversion. arXiv preprint arXiv:2208.01618, 2022.

[3]: Jascha Sohl-Dickstein, Eric A. Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. CoRR, abs/1503.03585, 2015.

[4]: https://huggingface.co/datasets/linoyts/3d_icon

Show and Tell: A Neural Image Caption Generator *Oriol Vinyals, Alexander Toshev, Samy Bengio, Dumitru Erhan CVPR 2015*

[5]: https://arxiv.org/abs/1411.4555

Show, Attend and Tell: Neural Image Caption Generation with Visual Attention *Kelvin Xuetal. ICML 2015*

[6]: *https://arxiv.org/abs/1502.03044*

Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering
Peter Anderson et al. CVPR 2018

[7]: https://arxiv.org/abs/1707.07998

Microsoft Common Objects in Context (COCO) 2017

Large-scale dataset with 330k images and 1.5 million captions.

[8]: https://cocodataset.org/#home

Flickr8k Dataset

[9]: https://forms.illinois.edu/sec/1713398

referenced in the paper *Hodosh, Mikolaj, Peter Young, and Julia Hockenmaier. "Framing image description as a ranking task: Data, models and evaluation metrics." JAIR, 2013.*

Github-Repository :

https://github.com/Insaf-Badri/IMAGE-TEXT-GENERATOR.git