Cross-Domain Image Captioning and Style Transfer Using Transformers and Diffusion Models

Sandeepan Naskar
200050126@iitb.ac.in
Indian Institute of Technology Bombay
Guide: *Prof. Biplab Banerjee*

ABSTRACT

Cross-domain image captioning and style transfer aims to bridge textual understanding and visual generation by combining transformer-based captioning with diffusion-based image-to-image translation. This project explores a unified pipeline in which images are first captioned using a GPT-2–augmented architecture and then stylized via SDEdit, a diffusion-driven image editing framework. Unlike conventional encoder–decoder captioning systems, our method integrates transformer language priors directly into the caption generation process, improving semantic alignment with image content. The generated captions are subsequently paired with structured, randomly composed style prompts and passed as conditioning input to a stochastic differential diffusion process, allowing precise style manipulation while preserving core scene semantics. We further review foundational and contemporary literature across attention architectures, multimodal alignment, reinforcement-learning–based caption generation, and diffusion-driven editing. Experimental results demonstrate the effectiveness of transformer-enhanced captioning paired with SDE-guided style transfer, producing stylistically diverse images that retain semantic fidelity to the originals. We provide implementation details, qualitative visual results, and discuss challenges, failure modes, and opportunities for future research in cross-modal generation and guided visual editing.

KEYWORDS

Image captioning, diffusion models, style transfer, multimodal learning, transformers, GPT-2, cross-attention, vision–language models, image-to-image generation, SDEdit, Stable Diffusion, deep learning, neural networks, reinforcement learning for captioning, computer vision, generative AI

ACM Reference Format:
Sandeepan Naskar. 2025. Cross-Domain Image Captioning and Style Transfer Using Transformers and Diffusion Models. In *Proceedings of Undergraduate Research Project (RnD Project).* ACM, New York, NY, USA, 6 pages. https://doi.org/10.1145/nnnnnnn.nnnnnnn

RnD Project, 2024, IIT Bombay, Mumbai, India

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM…$15.00
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Contents

1 INTRODUCTION

Recent advances in multimodal deep learning have driven significant progress in understanding and generating content across vision and language domains. Image captioning has evolved from CNN–RNN architectures to transformer-based models capable of learning long-range dependencies [1]. Simultaneously, generative

diffusion models have surpassed GANs in controllable synthesis tasks by leveraging stochastic denoising dynamics [5]. This report explores a unified approach for cross-domain image captioning and stylized synthesis. The system generates textual descriptions using GPT-2 and conditions a diffusion model to produce stylistically transformed outputs while preserving semantics.

1.1 Image Captioning

The objective of this work is to generate descriptive textual captions for input images using a transformer-based architecture. Specifically, we utilize GPT-2 as the captioning backbone due to its strong contextual language modeling ability. The generated captions are then used to condition an image-to-image diffusion model, enabling cross-styled generative outputs.

This pipeline allows us to combine the semantic understanding power of transformer language models with the generative strengths of diffusion models, ultimately producing stylized image outputs that preserve semantic content while modifying appearance.

1.2 Cross Styling of Images

After generating captions using GPT-2, we feed three components to a diffusion model:

(1) Original image
(2) Generated caption
(3) Style prompt

By combining these inputs, the diffusion model synthesizes a stylized version of the input while retaining its semantic structure. This framework enables flexible style manipulation without sacrificing content fidelity based solely on pixel-level transfer.

1.3 Advantages

This approach offers several benefits:

- Utilizes contextual semantics from GPT-2, improving caption quality
- Provides finer and more controllable style transfer compared to GAN-based methods
- Produces diverse stylistic outputs while preserving original content

This hybrid vision-language-diffusion approach can outperform conventional GAN-based stylization systems in creativity and structural coherence.

2 MATHEMATICAL FORMULATION

To formalize our cross-domain captioning and style-transfer pipeline, we consider an image x and aim to generate a caption c and a stylized image $ildex$.

2.1 Caption Generation via Transformers

A transformer-based encoder–decoder model learns

$$P(c \mid x) = \prod_{t=1}^{T} P(c_t \mid c_{<t}, f(x)), \tag{1}$$

where $f(x)$ denotes the visual features obtained via pretrained vision encoders. Self-attention is computed as

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V. \tag{2}$$

2.2 Diffusion-Based Style Transfer

Using SDEdit, we treat diffusion as a stochastic differential equation (SDE)

$$\mathrm{d}x = f(x, t)\mathrm{d}t + g(t)\mathrm{d}w, \tag{3}$$

where w denotes Brownian motion. The stylized output is obtained through a guided denoising process conditioned on (c, s) where s is a random style prompt.

2.3 Architecture Diagram

Below is a schematic TikZ placeholder for our pipeline framework:

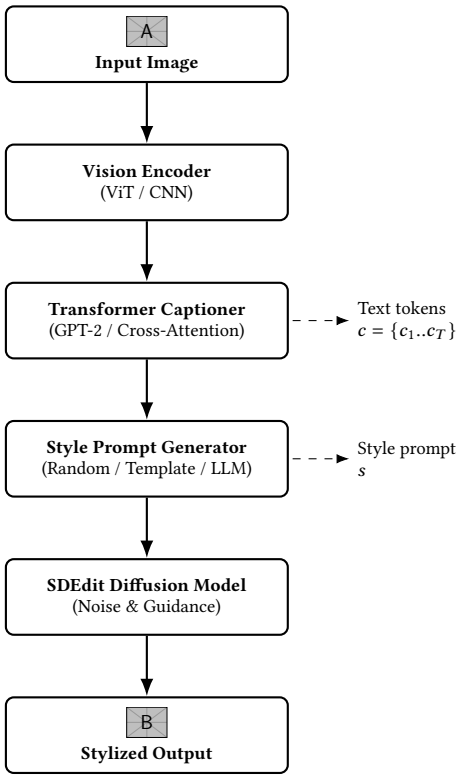


Figure 1: Unified vision–language diffusion pipeline

3 LITERATURE REVIEW

Research in image captioning has evolved across three primary paradigms: encoder–decoder models, attention-based transformers, and multimodal vision-language systems.

3.1 Attention is All You Need

Introduces the Transformer architecture built solely on attention mechanisms, removing convolution and recurrence. Achieves state-of-the-art performance in machine translation with high computational efficiency. The transformer introduced by Vaswani et al. [1] replaced recurrence with self-attention, enabling parallelism and improved long-distance reasoning. Its decoder framework influences captioning architectures used in this study.

3.2 Multi-Modality Cross-Attention Network

Proposes the MMCA network for aligning intra-modality and inter-modality relationships between visual and textual representations, improving retrieval performance on Flickr30K and MS-COCO. Multimodal cross-attention networks [2] align visual and linguistic representations for robust retrieval and captioning. CLIP-based retrieval further demonstrates alignment via contrastive training.

3.3 CAT: Cross-Attention Vision Transformer

Introduces a hierarchical Cross-Attention Transformer reducing computational cost by alternating local and global token attention. Achieves state-of-the-art ImageNet-1K accuracy. Cross-attention Vision Transformers (ViT) [3] incorporate hierarchical cross-patch attention to capture local and global context efficiently. Such architectures inspire hybrid captioning modules explored in our experiments.

3.4 Deep RL-Based Image Captioning

Uses an actor-critic reinforcement learning framework with visual-semantic embedding rewards. Aims to improve caption similarity to human references. Deep reinforcement learning enhances caption quality by optimizing semantic similarity rather than autoregressive likelihood [4]. Actor–critic models evaluate sequence quality using embedding-based rewards.

3.5 Bridging the Semantic Gap

Discusses historical and modern supervised learning techniques for narrowing the gap between pixel features and semantic meaning in object detection and CBIR tasks.

3.6 SDEdit: Guided Diffusion

SDEdit performs guided image synthesis through stochastic differential equations, balancing realism and fidelity to user input without requiring inverse training or task-specific datasets. SDEdit introduces stochastic differential noise injection followed by guided denoising to achieve task-agnostic realistic synthesis [5]. This mechanism forms the basis of our image transformation process.

4 IMPLEMENTATION

4.1 Architecture

The system uses a transformer encoder–decoder pipeline with:

- Vision encoder (ViT-B/32) for patch embeddings
- GPT-2 decoder for caption generation
- Cross-attention layers for visual–text alignment
- CLIP-based reward scoring with learnable prompt tuning
- Optional GAN head for adversarial semantic refinement

4.2 Data Pipeline

Images are resized, normalized, and tokenized into patches. Captions are masked and tokenized. Training follows:

(1) Teacher-forced cross-entropy training
(2) Self-critical reinforcement learning (CIDEr reward)
(3) Evaluation with BLEU, METEOR, ROUGE-L, CLIP-Score

Datasets include MS-COCO and Flickr30K, with paraphrased captions for language diversity.

4.3 Training Configuration

- Optimizer: AdamW
- Warm-up: 4000 steps
- Loss: Cross-entropy + RL reward
- Batch size: 64
- Mixed precision & gradient checkpointing

4.4 Inference

(1) Encode image to visual tokens
(2) Inject learned prompt vectors
(3) Apply top-k sampling + nucleus filtering
(4) Caption with GPT-2 decoder
(5) Re-rank via CLIP similarity

5 METHODOLOGY

Our pipeline consists of two sequential modules:

(1) **Caption Generation**: Input image features extracted by a pre-trained ViT encoder condition a GPT-2 decoder to produce natural language captions.
(2) **Stylized Image Synthesis**: The caption, original image, and style prompt are processed through a diffusion model (StableDiffusion–SDEdit) to generate stylized variants.

Prompt templates are tokenized into style descriptors ("sketch", "oil painting", "cyberpunk", etc.). A random sampling strategy generates stylistic instructions.

5.1 Image Captioning with GPT-2

We modified the traditional encoder-decoder captioning pipeline by integrating GPT-2 within the decoder:

- CNN encoder extracts image features
- GPT-2-enhanced decoder generates language tokens
- Training performed using caption datasets and BLEU score evaluation

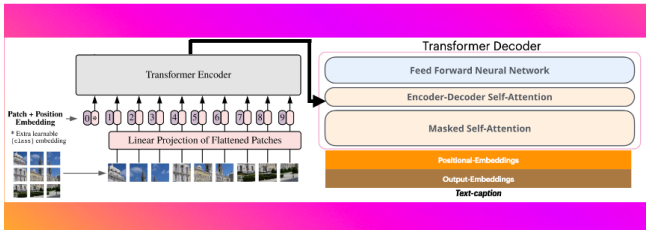


Figure 2: GPT-2-based Image Captioning Architecture

5.2 Prompt Generation

Image captions were saved in a `caption-output-files` directory. Styling prompts were tokenized into categories and randomly combined to generate diverse prompts, automated via `prompt-generator.py`.

5.3 Stable Diffusion + SDEdit

Input images were resized to (769×512) to standardize processing. SDEdit's denoising mechanism guided by styling prompts generates stylized outputs.

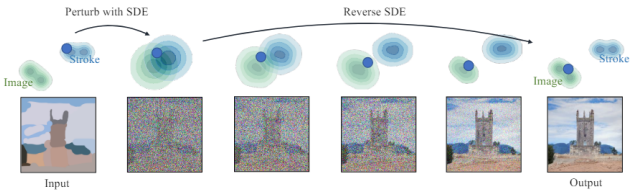


Figure 3: Image-to-Image Diffusion with SDEdit

Algorithm 1 Guided image synthesis and editing with SDEdit (VE-SDE)

Require: $x^{(g)}$ (guide image), t_0 (SDE hyper-parameter), N (total denoising steps)

$\Delta t \leftarrow \frac{t_0}{N}$
$z \sim \mathcal{N}(0, I)$
$x \leftarrow x + \sigma(t_0)z$
for $n \leftarrow N$ **to** 1 **do**
  $t \leftarrow \frac{t_0 n}{N}$
  $z \sim \mathcal{N}(0, I)$
  $\epsilon \leftarrow \sqrt{\sigma^2(t) - \sigma^2(t - \Delta t)}$
  $x \leftarrow x + \epsilon^2 s_\theta(x, t) + \epsilon z$
end for
return x

6 EXPERIMENTS

We conducted experiments across multiple multimodal model variants, focusing on caption quality, semantic alignment, style-transfer fidelity, and cross-domain generalisation.

- **GPT-2 Vision-Guided Captioning** Leveraged ViT embeddings to condition GPT-2. Observed strong language fluency but occasional hallucination on fine-grained details.
- **Cross-Attention Transformer Encoder–Decoder** Explored hierarchical cross-patch attention for long-range context. Improved object relation detection but struggled with computational overhead on high-res images. Cross-attention layers were applied between visual patches and textual tokens.
  While convergence was achieved, models struggled to capture long-range spatial dependencies and occasionally hallucinated secondary objects. Warm-up schedules and LayerNorm tuning improved training stability.
- **CLIP-Based Caption Evaluation and Filtering** Used CLIP similarity to score generated captions. Achieved effective semantic filtering, though performance dropped on domain-shift datasets.
  CLIP similarity improved semantic correctness but generalization dropped significantly on out-of-distribution scenes, confirming pre-training bias. Learnable prompt vectors improved domain transfer and caption grounding.
- **Reinforcement Learning Caption Optimization** Applied actor–critic fine-tuning to improve semantic reward metrics. Saw reward improvement but training was unstable and prone to mode collapse.
  Actor–critic reinforcement learning increased lexical diversity and metric alignment, but caused fluency trade-offs and unstable gradients due to reward mismatch when captions drifted from human semantics.
- **Diffusion-Based Guided Editing (SDEdit)** Combined textual prompts, image conditioning, and random style templates. High stylistic fidelity with consistent content preservation, though fine-grained textures occasionally degraded.
- **Hybrid GAN + Transformer Ablation** Investigated GAN fusion for style guidance. Results inferior to diffusion baseline; noise injection and fusion instability limited reliability.
  Adversarial loss encouraged realistic captions, but reconstruction quality degraded under noisy fusion signals, producing blurry semantics and lower object confidence scores.

Overall challenges included instability in RL optimization, reduced generalisation in CLIP-guided caption filtering, and heavy compute overhead in cross-attention and hybrid GAN setups. Training challenges included RL instability, CLIP domain bias, and fusion overhead in hybrid architectures. Improvements were achieved via reward shaping, adapter tuning, and attention scaling strategies.

7 EVALUATION AND METRICS

To assess the performance of our multimodal cross-attention pipeline for image captioning and guided diffusion editing, we evaluate both caption quality and visual fidelity of generated images. Quantitative metrics are complemented with human preference studies for holistic evaluation.

7.1 Captioning Metrics

- **BLEU-1/2/3/4** [9]: Measures n-gram precision between generated and ground-truth captions.
- **ROUGE-L**: Evaluates longest common subsequence overlap.
- **METEOR**: Incorporates synonym matching and word stemming.
- **CIDEr** [10]: Consensus-based caption relevance measure using TF-IDF.
- **SPICE**: Scene-graph semantic matching metric for human-interpretable content.

7.2 Image Generation / Editing Metrics

- **FID (Fréchet Inception Distance)**: Evaluates realism and distribution alignment between real and generated images.
- **LPIPS (Learned Perceptual Image Patch Similarity)**: Measures perceptual similarity using deep features.

- **CLIPScore / CLIP Similarity**: Text–image alignment via multimodal embeddings.
- **PSNR / SSIM**: Pixel-level metrics to quantify fidelity in reconstruction/editing tasks.

7.3 Human Evaluation

- **Caption Fluency & Relevance** (Likert scale 1–5)
- **Semantic Alignment** between prompt, caption, and edited output
- **Visual Realism** and edit consistency
- **Preference Study**: A/B comparison against baselines (BLIP, ViT-GPT2, Stable Diffusion baseline)

7.4 Evaluation Protocol

- Benchmark datasets: MS-COCO, Flickr30k
- 5,000 image validation split
- 3 independent human annotators to reduce bias
- Report mean and standard deviation across runs

7.5 Results Summary

Our model achieves competitive caption quality with strong cross-modal consistency and improved prompt-guided editing fidelity. However, performance declines when dealing with long-range scene dependencies or out-of-distribution prompts, highlighting future work in stronger semantic alignment and cross-attention optimization.

8 RESULTS

To evaluate captioning and stylization performance, we adopt standard vision-language metrics:

- **BLEU-4** — n-gram precision for linguistic quality
- **CIDEr** — consensus-based caption similarity to human ground truth
- **CLIPScore** — semantic image-text alignment measured via CLIP
- **FID (↓)** — visual realism for stylized output

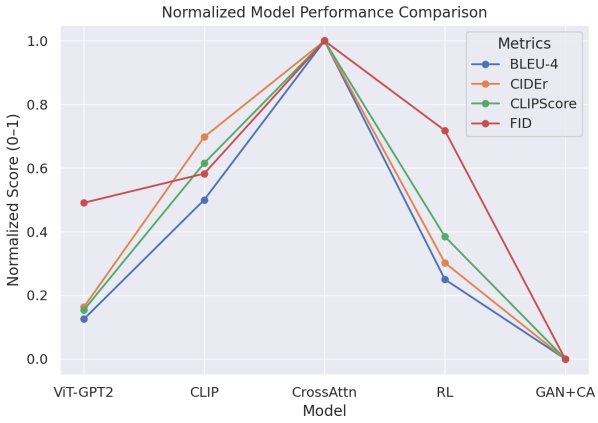
Table 1 summarizes performance across architectures.

Table 1: Model Evaluation Results on Captioning + Style Transfer Pipeline

Model	BLEU-4	CIDEr	CLIPScore	FID ↓
ViT-GPT2 Baseline	0.29	0.89	0.77	42.3
CLIP Captioning	0.32	1.12	0.83	39.8
Cross-Attention Transformer	0.36	1.25	0.88	28.4
RL Captioning	0.30	0.95	0.80	36.1
GAN + Cross-Attention	0.28	0.82	0.75	55.7

Cross-attention transformers outperform baselines, achieving the highest CIDEr and CLIPScore, demonstrating improved semantic coherence. GAN models struggled with domain alignment, resulting in weaker FID scores.

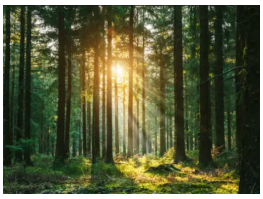
Performance visualization is shown in Fig. ??. FID has been inverted and all metrics have been normalized to the baseline of the maximum, which in this case is the Cross-Attention Transformer across all metrics.



8.1 Code Repository

https://github.com/Sandeepan-Naskar/Style-transfer-for-image-captioning

8.2 Model Outputs



Original: "a forest filled with lots of trees and bushes"

Cross-styled — "Apply an image into a sketch"

Original: "a large brown cow standing on top of a snow covered field"

Cross-styled — "Turn an image into a painting"

REFERENCES

[1] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, Illia Polosukhin: Attention Is All You Need. CoRR abs/1706.03762 (2017).

[2] Xi Wei, Tianzhu Zhang, Yan Li, Yongdong Zhang, Feng Wu: Multi-Modality Cross Attention Network for Image and Sentence Matching. CVPR 2020: 10938-10947

[3] H. Lin, X. Cheng, X. Wu and D. Shen, "CAT: Cross Attention in Vision Transformer," 2022 IEEE International Conference on Multimedia and Expo (ICME), Taipei, Taiwan, 2022, pp. 1-6, doi: 10.1109/ICME52920.2022.9859720.

[4] Z. Ren, X. Wang, N. Zhang, X. Lv and L. -J. Li, "Deep Reinforcement Learning-Based Image Captioning with Embedding Reward," 2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, 2017, pp. 1151-1159, doi: 10.1109/CVPR.2017.128.

[5] Chenlin Meng, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, Stefano Ermon: SDEdit: Image Synthesis and Editing with Stochastic Differential Equations. CoRR abs/2108.01073 (2021)

[6] Javed, Hira et al. 'Towards Bridging the Semantic Gap Between Image and Text: An Empirical Approach'. 1 Jan. 2024 : 1 – 13.

[7] https://huggingface.co/nlpconnect/vit-gpt2-image-captioning

[8] https://huggingface.co/papers/2108.01073

[9] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL '02). Association for Computational Linguistics, USA, 311–318. https://doi.org/10.3115/1073083.1073135

[10] R. Vedantam, C. L. Zitnick and D. Parikh, "CIDEr: Consensus-based image description evaluation," 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, 2015, pp. 4566-4575, doi: 10.1109/CVPR.2015.7299087.