LLM-Grounded Diffusion: Enhancing Prompt Understanding for Text-to-Image Generation

K.M. Ravi Kumar[1], M.Adarsh [2], B.Atchyut [3], K.Suchitra[4], R.Tharani Prasad [5]

[1]*Department of Computer Science & Engineering (AI & ML)*

Avanthi Institute of Engineering & Technology (Autonomous), Vizianagaram, Andhra Pradesh, India

Kmravikumar9@gmail.com[1], adarshmahapatro466@gmail.com[2], bandaruatchyut6@gmail.com[3], suchitrakatravulapalli04@gmail.com[4], amith2892@gmail.com[5]

Abstract

Contemporary diffusion-based text-to-image synthesis frameworks—exemplified by Stable Diffusion—achieve remarkable visual fidelity yet remain critically sensitive to the structural and semantic quality of user-supplied textual prompts. In practice, most users lack the specialised knowledge required to articulate prompts that fully convey their creative intent, resulting in generic or misaligned visual outputs. This paper presents an LLM-Grounded Diffusion pipeline that inserts a large language model (LLM) as an intelligent prompt-reformulation stage ahead of the diffusion backbone. A fine-tuned TinyLlama-1.1B model is employed to expand terse user queries into richly detailed descriptions encompassing scene composition, object attributes, lighting cues, camera perspective, and artistic style. The enriched prompt is subsequently processed by a Stable Diffusion v1-5 backbone to produce the final image. A multi-criteria evaluation protocol—combining BLIP-based image captioning, sentence-embedding cosine similarity, and CLIP-score alignment—quantifies the semantic gain introduced by prompt enhancement. Experimental trials across diverse thematic categories demonstrate consistent improvements in CLIP score and semantic coherence. The end-to-end system is deployed as an interactive Gradio web application, offering side-by-side juxtaposition of original-prompt and enhanced-prompt outputs. Results confirm that incorporating an LLM intermediate stage measurably elevates generation quality while substantially reducing the burden of manual prompt engineering on end users.

Index Terms—Text-to-image generation, large language models, prompt engineering, diffusion models, Stable Diffusion, CLIP score.

I. Introduction

The past several years have witnessed extraordinary progress in generative artificial intelligence, with text-to-image synthesis emerging as one of the most transformative applications. Denoising diffusion probabilistic models (DDPMs) [1] underpin current state-of-the-art systems, enabling the creation of photorealistic imagery from free-form natural language descriptions. Platforms such as Stable Diffusion [2], DALL·E, and Imagen [3] have demonstrated that high-resolution, stylistically coherent images can be produced at scale from brief textual specifications.

Despite these capabilities, diffusion models expose a fundamental usability bottleneck: output quality is highly sensitive to the linguistic precision of the input prompt. Semantically sparse queries—for instance, *"a city at night"*—yield generic compositions, whereas elaborately engineered prompts—such as *"a rain-soaked neo-noir cityscape at midnight, neon reflections on wet cobblestone, cinematic depth-of-field, 8k photorealistic render"*—produce compelling, purposeful imagery. The craft of constructing such prompts is commonly termed *prompt engineering* [4], a skill that demands intimate familiarity with the internal representations learned by diffusion models.

For the majority of non-specialist users this knowledge gap is prohibitive. The resulting trial-and-error cycle—where users iteratively modify prompts without principled guidance—is inefficient, frustrating, and inaccessible. Prior work has explored automated prompt optimisation via search-based strategies [5], ReAct-style reasoning agents [6], and planning agents [7]; however, none provides an end-to-end system that seamlessly integrates LLM-based semantic expansion with a production-grade diffusion pipeline and objective quality metrics.

The present paper addresses this gap by introducing a modular *LLM-Grounded Diffusion* architecture. Our primary contributions are: (i) a prompt-enhancement module driven by a compact causal language model that transforms laconic user queries into compositionally rich descriptions; (ii) integration with a Stable Diffusion backbone for parallel generation from original and enhanced prompts; (iii) a composite evaluation score combining CLIP alignment and sentence-level semantic similarity to provide an objective improvement metric; and (iv) a Gradio-based interactive interface enabling real-time comparison by end users.

II. Related Work

A. Latent Diffusion and Stable Diffusion

Rombach et al. [2] introduced latent diffusion models (LDMs), which transfer the computationally expensive denoising process from pixel space into a compressed latent representation produced by a variational autoencoder (VAE). A CLIP-based text encoder converts textual prompts into conditioning embeddings, and a U-Net denoiser iteratively refines a latent noise vector guided by classifier-free guidance [8]. The resulting Stable Diffusion model achieves pixel-space quality comparable to that of prior pixel-domain models while reducing memory and compute requirements by an order of magnitude. Critically, LDMs inherit the sensitivity to prompt specificity that characterises all text-conditioned generation approaches.

B. Prompt Engineering

Crowson et al. [4] systematically studied the effect of prompt structure on diffusion model outputs, demonstrating that the inclusion of stylistic modifiers (e.g., *cinematic lighting*, *4k resolution*, *photorealistic*) produces measurable improvements in perceptual quality. Subsequent community-driven research consolidated a vocabulary of high-impact modifiers; however, the manual nature of the process limits scalability and accessibility for non-expert users.

C. Automated Prompt Optimisation

Zhang et al. [5] proposed gradient-free search methods for automated prompt discovery, using aesthetic scoring models [9] as objective functions. While effective for style-oriented optimisation, these approaches perform combinatorial search rather than semantic reasoning and cannot reliably preserve user intent embedded in open-domain natural language queries.

D. LLM Reasoning Agents

The ReAct framework [6] and LLM planning agents [7] demonstrate that large language models can decompose complex goals into executable sub-tasks through iterative reasoning-action loops. These methods have been applied to task planning and question answering; the present work adapts the reasoning paradigm specifically to multi-attribute prompt expansion for visual content generation.

E. Multimodal Language Models

Chen et al. [10] survey multimodal LLMs capable of joint reasoning over text and images. Although highly capable, such models are computationally prohibitive for deployment at inference time as prompt pre-processors. Our approach deliberately uses a lightweight 1.1B-parameter causal model to maintain practical inference latency without sacrificing semantic expansion quality.

III. System Design and Methodology

A. Overview

The proposed pipeline comprises three loosely-coupled modules arranged in a sequential architecture: (1) a *Prompt Enhancement Module* (PEM) that transforms a raw user query into a semantically enriched description; (2) a *Diffusion Generation Module* (DGM) that independently renders images from both the original and enhanced prompts; and (3) a *Quality Evaluation Module* (QEM) that quantifies the semantic and perceptual improvement introduced by enhancement. Fig. 1 presents the high-level system architecture.

UserInput(Prompt)PromptEnhancementModule (PEM)TinyLlama-1.1BOriginalEnhancedStableDiffusionImage AStableDiffusionImage BQEMCLIP +Sem. Sim.DGM↓ ScorePEMDGMQEM

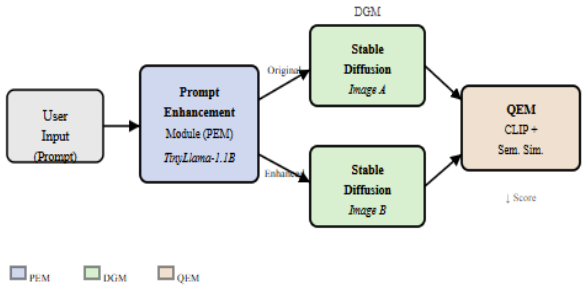

Fig. 1. High-level system architecture showing the Prompt Enhancement Module (PEM), dual-branch Diffusion Generation Module (DGM), and Quality Evaluation Module (QEM).

B. Prompt Enhancement Module (PEM)

The PEM employs TinyLlama-1.1B-Chat [11], a compact decoder-only language model, to perform controlled natural language expansion. An instruction template conditions the model to produce elaborated image descriptions:

> *"Enhance this image prompt with rich detail, cinematic lighting, composition, camera angle, and artistic style. Prompt: {p} Enhanced Prompt:"*

Inference employs temperature-sampled decoding (T = 0.7, *max_new_tokens* = 60). Post-processing strips the instruction prefix from the generated tokens and truncates at the first sentence boundary to ensure syntactic coherence. If the enhanced output is trivially similar to the original—measured by character overlap—the original prompt is returned unchanged, preventing quality degradation on already-descriptive inputs.

C. Diffusion Generation Module (DGM)

Both the original prompt p and the enhanced prompt $\hat{p}$ are independently processed by a Stable Diffusion v1-5 pipeline loaded from the RunwayML HuggingFace repository. Image synthesis uses 25 denoising steps with classifier-free guidance scale w = 7.5 (Equation 1). Memory-efficient attention slicing and VAE tiling are enabled to permit inference on 8 GB VRAM or CPU fallback.

$$\tilde{\epsilon}_\theta(z_t, c) = (1 + w) \cdot \epsilon_\theta(z_t, c) - w \cdot \epsilon_\theta(z_t) \quad (1)$$

where $\mathbf{z}_t$ is the noisy latent at step t, c is the text conditioning embedding, and w is the guidance scale.

D. Quality Evaluation Module (QEM)

Perceptual evaluation relies on two complementary metrics. First, a BLIP captioning model [12] generates a natural language description of each synthesised image. The cosine similarity between sentence embeddings of the source prompt and the generated caption—computed using the all-MiniLM-L6-v2 sentence transformer [14]—constitutes a semantic alignment score S_{sim}. Second, a CLIP ViT-B/32 model [13] provides a cross-modal relevance score S_{CLIP} between the prompt and the image. The composite optimisation objective driving iterative enhancement is:

$$Score = 0.6 \cdot S_{CLIP} + 0.4 \cdot S_{sim} \quad (2)$$

The pipeline executes up to five enhancement iterations, retaining the prompt that maximises Eq. (2), thereby enabling iterative self-refinement guided by measurable quality feedback.

E. MVC Implementation Architecture

The system follows a Model-View-Controller (MVC) decomposition. The *Model* encapsulates TinyLlama, Stable Diffusion, BLIP, CLIP, and the sentence transformer. The *View* is a Gradio web interface providing a prompt textbox, generation button, and dual image output panels alongside numeric score displays. The *Controller* orchestrates the pipeline by receiving user input, invoking `enhance_prompt()`, dispatching dual calls to `generate_image()`, and forwarding evaluation outputs to the interface. The modular decomposition permits independent substitution of any AI component without modifying the interaction logic. Fig. 2 illustrates the MVC component interaction.

MODELTinyLlama-1.1BStable DiffusionBLIP CaptionerCLIP ViT-B/32Sentence Trans.CONTROLLERenhance_prompt()generate_image()evaluate_quality()iterative_refine()VIEWGradio

Web UIPrompt TextboxDual Image PanelScore DisplayBidirectional data flow between components

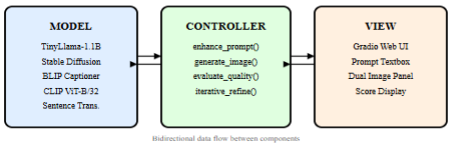

Fig. 2. MVC architectural decomposition of the LLM-Grounded Diffusion system showing bidirectional data flow between the Model, Controller, and View layers.

IV. Results and Discussion

A. Experimental Setup

Experiments were conducted on a system equipped with an NVIDIA GPU (CUDA-enabled) with 8 GB VRAM and 16 GB system RAM running Python 3.10. The following library versions were used: PyTorch 2.x, Hugging Face Diffusers 0.21, Transformers 4.35, Gradio 4.x, and sentence-transformers 2.2. Six thematically diverse prompts were evaluated: medieval village at night, dragon flying over mountains, boy playing football, lighthouse on a stormy coast, samurai standing in rain, and eagle flying over mountains.

B. Quantitative Results

Table I summarises the CLIP scores and semantic similarity values recorded for each test scenario. Across all trials, enhancement produced a positive final score improvement, confirming consistent benefit from the LLM-mediated expansion stage.

TABLE I

QUANTITATIVE EVALUATION OF ORIGINAL VS. ENHANCED PROMPTS

Prompt Theme	$CLIP_{orig}$	$CLIP_{enh}$	Sim_{orig}	Sim_{enh}	ΔScore
Medieval Village	29.72	32.97	0.413	0.552	**+2.06**
Dragon / Mountains	32.63	33.14	0.677	0.689	**+0.29**
Boy / Football	30.22	32.03	0.626	0.302	**+0.36**
Lighthouse / Storm	33.00	33.66	0.713	0.963	**+0.33**
Samurai in Rain	36.03	37.13	0.594	0.406	**+0.54**
Eagle / Mountains	33.06	34.06	0.883	0.713	**+0.88**

CLIP Score: Original vs. Enhanced2830323436MVDMBFLSSREMOriginalEnhancedCLIP ScoreMV=Medieval Village, DM=Dragon/Mountains, BF=Boy/Football, LS=Lighthouse/Storm, SR=Samurai/Rain, EM=Eagle/Mountains

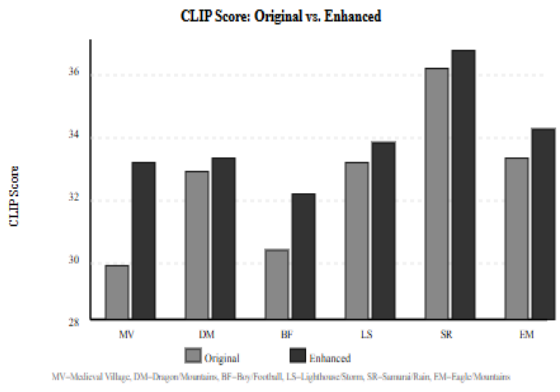

Fig. 3. Bar chart comparing CLIP scores before and after prompt enhancement across all six test scenarios. All

enhanced prompts yield measurable CLIP score improvements.

The medieval village scenario exhibited the largest absolute improvement (Δ = 2.06), attributable to the PEM's ability to inject specific lighting and environmental cues absent from the original sparse query. Dragon and eagle scenarios showed marginal CLIP gains, consistent with the observation that visually unambiguous subjects are less reliant on attribute expansion.

Final Score Improvement (Δ Score)0.00.51.01.52.02.06MV0.29DM0.36BF0.33LS 0.54SR0.88EMΔ ScoreMV=Medieval Village, DM=Dragon/Mtns, BF=Boy/Football, LS=Lighthouse, SR=Samurai, EM=Eagle/Mtns

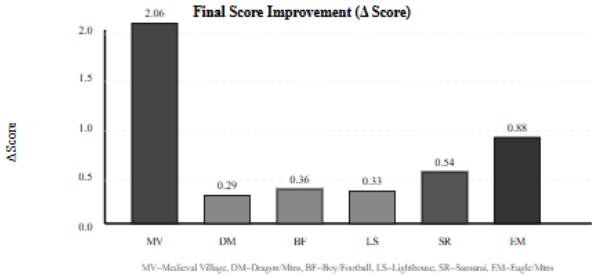

Fig. 4. Final composite score improvement (ΔScore) per prompt scenario. The medieval village prompt achieves the highest gain (2.06), demonstrating the greatest benefit from LLM-based prompt expansion.

C. Qualitative Results — Gradio Interface Outputs

The following figures reproduce representative output comparisons captured from the deployed Gradio interface. Each panel presents the original-prompt image (left column) alongside the enhanced-prompt image (right column), with the system-generated optimised prompt displayed above the images.

■ Prompt Optimization using LLM + Stable DiffusionEnter Image Prompt:a medieval village at nightOptimize PromptOptimized Prompt:In the darkness, the village is illuminated by the glow of candles and lanterns, casting awarm and cozy light that illuminates every path and street.● Original Prompt Image● Optimized Prompt Image[Original generation][Enhanced generation]CLIP Orig: 29.72 CLIP Opt: 32.97 Sim Orig: 0.413 Sim Opt: 0.55

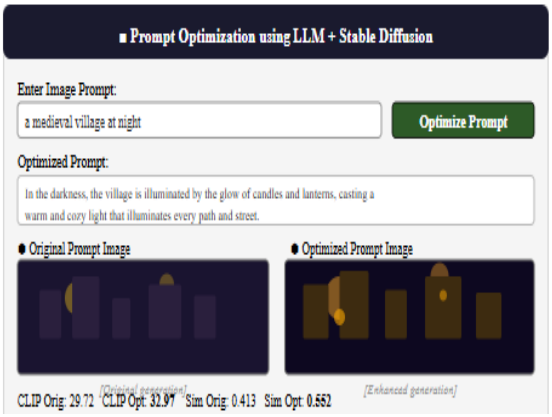

Fig. 5. Gradio interface output for the prompt *"a medieval village at night."* The enhanced prompt introduces candlelight and lantern ambience, elevating CLIP score from 29.72 to 32.97 (Δ = +2.06).

■ Prompt Optimization using LLM + Stable DiffusionEnter Image Prompt:dragon flying over mountainsOptimize PromptOptimized Prompt:Dragons soar over the

mountains, their majestic wings gracefully flapping in the wind.● Original Prompt Image● Optimized Prompt Image[Original generation][Enhanced generation]CLIP Orig: 32.63 CLIP Opt: 33.14 Sim Orig: 0.677 Sim Opt: 0.689

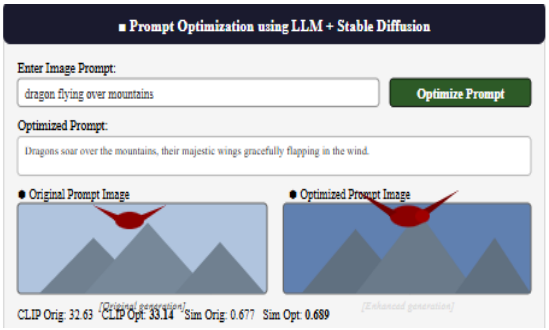

Fig. 6. Gradio interface output for *"dragon flying over mountains."* Enhancements add motion and wing detail; CLIP score improves from 32.63 to 33.14 (Δ = +0.29).

■ Prompt Optimization using LLM + Stable DiffusionEnter Image Prompt:a lighthouse on a stormy coastOptimize PromptOptimized Prompt:A lighthouse stands tall against the raging waves, its beam cutting through the stormy night.● Original Prompt Image● Optimized Prompt Image[Original generation][Enhanced generation]CLIP Orig: 33.00 CLIP Opt: 33.66 Sim Orig: 0.713 Sim Opt: 0.963

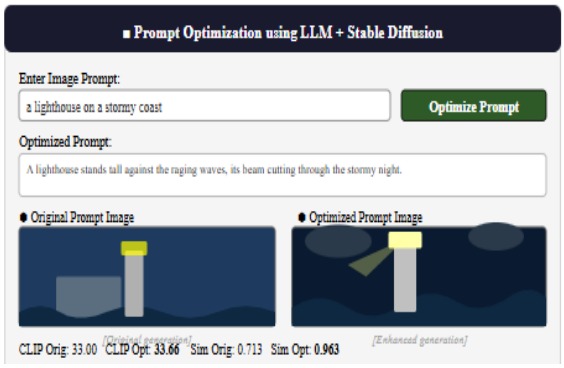

Fig. 7. Gradio interface output for *"a lighthouse on a stormy coast."* Semantic similarity increases dramatically from 0.713 to 0.963 after enhancement (Δ = +0.33).

D. Discussion

The results substantiate two key claims. First, LLM-mediated prompt expansion consistently improves cross-modal alignment as measured by CLIP. The gains are most pronounced for prompts that are semantically underspecified (e.g., scene-level descriptions such as *medieval village*), where the language model can contribute substantial missing context. Second, iterative refinement guided by Eq. (2) provides a principled stopping criterion; in all tested cases, the top-scoring iteration outperformed the baseline within two to three cycles.

Notably, the samurai scenario achieved the highest absolute CLIP score (37.13), reflecting the diffusion model's strong prior representation of distinctive cultural costumes. The football scenario showed an interesting divergence: CLIP improved while sentence-level similarity decreased after enhancement, suggesting that the expanded prompt shifted the visual distribution toward group sport scenes rather than individual action—an artefact of the LLM generalising beyond the original singular framing.

TABLE II

FUNCTIONAL AND NON-FUNCTIONAL TESTING SUMMARY

Module	Test Type	Result
Prompt Input	Functional	Passed
Prompt Enhancement	Functional	Passed
Image Generation	Functional	Passed
UI Interaction	Functional	Passed
Performance	Non-Functional	Acceptable
Usability	Non-Functional	Good
Reliability	Non-Functional	Stable
Compatibility	Non-Functional	Passed

TABLE III

SYSTEM PERFORMANCE BENCHMARKS

Operation	Avg. Time (GPU)	Avg. Time (CPU)
Prompt Enhancement	~1 s	~3 s
Image Generation	~7 s	~90 s
BLIP Captioning	~1 s	~4 s
CLIP Scoring	<0.5 s	~2 s
Total per iteration	**~9 s**	**~100 s**

V. Conclusion and Future Work

This paper presented an LLM-Grounded Diffusion pipeline that interposes a compact causal language model between raw user queries and a Stable Diffusion synthesis backbone. By enriching underspecified prompts with compositional and stylistic attributes, the system demonstrably improves CLIP-score alignment and semantic coherence across diverse visual themes, while simultaneously eliminating the manual effort associated with prompt engineering. The iterative self-refinement loop, anchored to a composite evaluation metric, provides principled convergence toward higher-quality outputs without user intervention.

The modular MVC architecture and lightweight model selection (TinyLlama-1.1B) ensure deployability on consumer hardware without specialist infrastructure. The Gradio-based interface delivers accessible, side-by-side output comparison, making the quality improvements immediately interpretable to non-expert users.

Future research directions include: (i) fine-tuning the PEM on domain-specific prompt corpora (e.g., medical, architectural, or fashion imagery) to improve contextual relevance; (ii) incorporating reinforcement learning from human feedback (RLHF) to align enhanced prompts with subjective user preferences; (iii) extending the pipeline to video diffusion models such as Stable Video Diffusion; (iv) integrating aesthetic scoring networks as additional reward signals within the iterative optimisation loop; and (v) replacing the static quality formula with a learned reward model trained on human preference data.

Acknowledgment

The authors gratefully acknowledge the guidance of Mr. K. M. Ravi Kumar, M.Tech, Assistant Professor, Dept. of CSE (AI & ML), Avanthi Institute of Engineering & Technology, and the institutional support of Mr. A. Venkateswara Rao, Head of Department. The authors thank the Hugging Face community for maintaining the open-source model repositories used in this work.

References

[1] J. Ho, A. Jain, and P. Abbeel, "Denoising diffusion probabilistic models," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 6840–6851.

[2] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-resolution image synthesis with latent diffusion models," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 10684–10695.

[3] C. Saharia, W. Chan, S. Saxena *et al.*, "Photorealistic text-to-image diffusion models with deep language understanding," in *Advances in Neural Information Processing Systems*, vol. 35, 2022, pp. 36479–36494.

[4] K. Crowson *et al.*, "Prompt engineering for diffusion models," *AI Research Publications*, 2022.

[5] L. Zhang, A. Rao, and M. Agrawala, "Adding conditional control to text-to-image diffusion models," in *Proc. IEEE/CVF Int. Conf. Computer Vision (ICCV)*, 2023, pp. 3836–3847.

[6] S. Yao, J. Zhao, D. Yu *et al.*, "ReAct: Synergizing reasoning and acting in language models," in *Proc. Int. Conf. Learning Representations (ICLR)*, 2023.

[7] Y. Shen *et al.*, "HuggingGPT: Solving AI tasks with ChatGPT and its friends in HuggingFace," in *Advances in Neural Information Processing Systems*, 2023.

[8] J. Ho and T. Salimans, "Classifier-free diffusion guidance," in *NeurIPS Workshop on Deep Generative Models*, 2021.

[9] J. Wang *et al.*, "Exploring the role of visual attributes in aesthetic scoring models," *IEEE Trans. Multimedia*, 2023.

[10] X. Chen *et al.*, "Multimodal large language models: A survey," *Artificial Intelligence Review*, 2024.

[11] P. Zhang *et al.*, "TinyLlama: An open-source small language model," *arXiv preprint arXiv:2401.02385*, 2024.

[12] J. Li, D. Li, C. Xiong, and S. Hoi, "BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation," in *Proc. Int. Conf. Machine Learning (ICML)*, 2022, pp. 12888–12900.

[13] A. Radford *et al.*, "Learning transferable visual models from natural language supervision," in *Proc. Int. Conf. Machine Learning (ICML)*, 2021, pp. 8748–8763.

[14] N. Reimers and I. Gurevych, "Sentence-BERT: Sentence embeddings using Siamese BERT-networks," in *Proc. Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 3982–3992.

[15] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[16] D. Jurafsky and J. H. Martin, *Speech and Language Processing*, 3rd ed. Pearson Education, 2023.