A. Research Proposal Summary

1.1 Title of research project:	Deep Generative Modeling for Automated Image Manipulation by Interpreting Text-Guided Prompts with Natural Language Instructions	
1.2. Keywords to identify reviewers:	Diffusion Trajectory, Image Manipulation, Latent Representation, Negative Prompting, Noise Vector, Null-Text Inversion, Prompt Engineering, Text-Guided Diffusion, Text Embeddings	
1.3. Research Focus Area:	1	
1.4. Principal Investigator:	Er. Dinesh Baniya Kshatri	dinesh@ioe.edu.np
1.5. Participating Faculties	Dr. Madhav Ghimire (Central Department of Physics)	madhav.ghimire@cdp.tu.edu.np
1.6. Participating Undergraduate Students	Abhinav Chalise	chalisezabhinav@gmail.com
	Nimesh Gopal Pradhan	nimeshgpradhan@gmail.com
	Nishan Khanal	nishan.077bei027@tcioe.edu.np
	Prashant Raj Bista	prashant.077bei032@tcioe.edu.np

B. Submission Checklist

1.1. Detail Research Proposal	1. Template guidelines have been followed ☑ 2. Research Proposal in PDF format has been submitted ☑
1.2. Proof of current affiliation with engineering campuses or colleges in Nepal as faculties or students	1. Proof of affiliation for faculty has been submitted (Appendix-1) ☑ 2. Proof of affiliation for undergraduate student has been submitted (Appendix - 2) ☑
1.3. Copy of certificate of highest education degree for faculties	1. Certificate of highest education degree for principal investigator has been submitted (Appendix - 3) ☑ 2. Equivalence of certificate from Tribhuvan University has been submitted (Appendix - 4) ☑
1.4. Scanned copy of declaration page including the signature of the principal investigator (available at the end of of this template)	1. Scanned copy of declaration page signed by the principal investigator has been submitted (Appendix -5) ☑

C. Title of Research Proposal

Deep Generative Modeling for Automated Image Manipulation by Interpreting Text-Guided Prompts with Natural Language Instructions

D. Description of Proposed Research

D.1: Research Abstract

This research delves into the complexities of enabling generative diffusion models to interpret and execute natural language instructions for automated image editing. Achieving precise semantic alignment between textual prompts and the resultant edited images necessitates overcoming linguistic nuances and image representations which stand as the primary challenge. Simultaneously, developing generative models proficient in executing intricate edits and generating high-fidelity images is also a sophisticated challenge. Overcoming these challenges holds the promise of revolutionizing image manipulation techniques across multiple domains.

The innovative aspect of our proposed solution lies in its ability to enable precise multi-object editing within a single prompt and a seamless undo mechanism devoid of residual artifacts. A layered editing method structured to optimize the manipulation of original images will facilitate the swift reversibility of applied edits. Our approach entails enabling multi-object editing by parsing multiple subjects within an image and the textual prompts, segmenting them, executing parallel edits, and merging the segments to produce the final image. Additionally, we're pursuing an iterative refinement approach, fine-tuning output images in each step by integrating user feedback.

The method to implement the proposed solutions integrates established models, such as Stable Diffusion, to process images through encoders, translating them into latent spaces conducive to granular manipulations. Leveraging well-established NLP models, textual prompts undergo segmentation, thereby creating a pipeline for iterative layered editing. By harnessing the latent space representations and existing NLP frameworks, this approach enables a seamless interplay between image and text, empowering users with precise, multi-object editing capabilities within a single prompt-driven framework. The latent representation of a layer of edited images is stored in a state management system enabling the functionality of redo/undo and negative insertions.

Recent advancements in text-guided image synthesis have focused on diffusion models for image manipulation, with methods like Prompt-to-Prompt and Negative Prompt Inversion demonstrating rapid, text-based image editing, albeit with limitations in handling complex compositions and accurate reconstruction of human images. Techniques like Prompt Tuning Inversion and Iterative Multi-granular Image Editor (EMILIE) further enhance image fidelity and control, though they face challenges in real-time applications and managing negative edit instructions, pointing to the need for more nuanced feature disentanglement in future research.

These advancements exhibit the potential to streamline content creation workflows, enrich human-AI interaction, and promote ethical image editing practices across multiple sectors. The democratization of image manipulation tools, catering to users with varying technical proficiencies, stands to significantly impact education and accessibility tools, improving creativity and efficiency.

D.2: Background and Review of Relevant Prior Work

In the accelerating field of deep generative modeling, automated image manipulation guided by textual prompts represents a groundbreaking intersection of computer vision and natural language processing which is gaining attention due to its wide-ranging applications. The ability to seamlessly enhance or

modify images through automated processes guided by textual prompts holds much significance. This research helps businesses, industries interact with visual data. The potential application of this research extends to various domains like graphic design, advertising, social media, and content creation. It can transform creative processes in animation studios, enabling artists to rapidly generate and edit concept art from text descriptions. For visual content creators and graphic designers, it offers a tool to prototype ideas and explore new designs efficiently. In the film industry, it aids in visualizing scenes, streamlines pre-production, and enhances visual effects, while photographers can leverage it to correct or enhance images By providing users the ability to edit through simple text prompts makes it more accessible to users of different technical skills.

Text-guided image synthesis and manipulation have become increasingly popular in recent years, especially with the advancement of Diffusion models. These models have been used in many research works for image manipulation. Prompt-to-Prompt [1], which achieved text-guided image editing without model refinement utilized cross-attention maps for controlled editing of images, primarily using the Imagen model, though it is adaptable to other models with cross-attention layers. Despite its innovative approach, Prompt-to-Prompt faced limitations, notably visible distortions and the challenge for users to generate suitable prompts for complex compositions. Prompt Tuning Inversion [2], introduced an innovative inversion method where the input image's information is encoded into a conditional embedding and incorporated learnable embedding in the sampling process. Utilizing the Latent Diffusion Model, Stable Diffusion [3], this method achieved impressive PSNR and SSIM scores, indicating high-quality editing while maintaining fidelity to the original image. However, it encountered difficulties when dealing with multiple objects in the input image. Null-text inversion [4], utilizing a pivotal inversion method, allowed for more efficient localized inversion. Null-text inversion method modified the unconditional textual embedding used for classifier-free guidance, enabling prompt-based editing without the need for model parameter tuning. While Null-text inversion is efficient, with inversion taking about a minute and an additional 10 seconds per edit, it is not suitable for real-time applications and tends to produce artifacts when dealing with human faces. Negative Prompt Inversion [5] presented a breakthrough with its method for rapid reconstruction of real images using Stable Diffusion. This approach, significantly faster than existing methods about 30 fold faster, also enabled prompt-to-prompt based quick real image editing. However, it struggled with accurately reconstructing images of people, despite its otherwise equivalent visual quality and speed benefits. EMILIE [6] emerged as a novel solution, introducing a latent iteration framework to reduce noise and artifacts by operating in the latent space. It integrated multi-granular control through denoising modulation and selective gradient updates for spatial editing, all without the need for retraining existing diffusion models. Yet, during experimentation, EMILIE revealed a critical limitation in handling negative edit instructions, like adding and subsequently removing sunglasses. This limitation highlights the necessity for future research in disentangling feature representations for each edit.

In current image editing methodologies, while users can change backgrounds, and styles, add objects, adjust resolutions, and apply iterative edits, these methods lack the capacity for precise "undo" operations and fail to edit multiple objects with a single prompt. Iterative multi-granular approaches, though effective in sequential editing, struggle with negative insertions and often leave artifacts even after an undo operation. Our research aims to address these gaps by developing a solution that enables effective multi-object editing and a cleaner undo functionality. This will involve refining the editing algorithms to identify and manipulate multiple elements within an image seamlessly and improve the undo process to

eliminate residual artifacts, thereby enhancing the overall precision and usability of automated image editing tools.

The research aims to develop a method that allows for precise editing of multiple objects within a single prompt and facilitates undoing previous edits without leaving artifacts. To perform multiple object editing we plan to detect the objects that prompt intends to edit by use of NLP and segment the image to detect multiple objects within it and pass each segment through the diffusion model guided by prompt and later combine these segments. We will also take an iterative approach to fine-tune the image to the desired outcome and use feedback to make adjustments in each step. For incorporating an undo mechanism we plan to use a layered approach where each edit is applied on a new layer, enabling easy removal of specific edits without affecting others. This approach will address the specific limitations of existing methods, which currently struggle with multi-object editing and cannot effectively reverse edits. By enabling more granular control and the ability to seamlessly undo changes, this solution could significantly enhance the flexibility and user-friendliness of automated image manipulation tools. It opens up new possibilities in creative image editing and can be particularly useful in professional settings like graphic design, where iterative editing and revision are common.

References

[1] A. Hertz, R. Mokady, J. Tenenbaum, K. Aberman, Y. Pritch, and D. Cohen-Or, "Prompt-to-Prompt Image Editing with Cross Attention Control," arXiv.org, Aug. 02, 2022. https://arxiv.org/abs/2208.01626 (accessed Nov. 19, 2023).

[2] W. Dong, S. Xue, X. Duan, and S. Han, "Prompt Tuning Inversion for Text-Driven Image Editing Using Diffusion Models," arXiv.org, May 08, 2023. https://arxiv.org/abs/2305.04441 (accessed Nov. 19, 2023).

[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution image synthesis with latent diffusion models," in 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Jun. 2022. Accessed: Nov. 22, 2023. [Online]. Available: http://dx.doi.org/10.1109/cvpr52688.2022.01042

[4] R. Mokady, A. Hertz, K. Aberman, Y. Pritch, and D. Cohen-Or, "Null-text Inversion for Editing Real Images using Guided Diffusion Models," arXiv.org, Nov. 17, 2022. https://arxiv.org/abs/2211.09794 (accessed Nov. 19, 2023).

[5] D. Miyake, A. Iohara, Y. Saito, and T. Tanaka, "Negative-prompt Inversion: Fast Image Inversion for Editing with Text-guided Diffusion Models," arXiv.org, May 26, 2023. https://arxiv.org/abs/2305.16807 (accessed Nov. 19, 2023).

[6] K. J. Joseph et al., "Iterative Multi-granular Image Editing using Diffusion Models," arXiv.org, Sep. 01, 2023. https://arxiv.org/abs/2309.00613 (accessed Nov. 19, 2023).

D.3: General Objective of the Research Project

The main goal of this research project is to advance the field of image editing by using text prompts, aiming to enhance and elevate the capabilities of existing models. The primary focus lies in developing novel algorithms for editing both real and synthetic images guided by text prompts. Building upon previous works, this research seeks to address limitations observed in the previous models like failure to perform multi-object editing and undo/redo operation. A key innovation involves the introduction of a state management system, designed to facilitate the undo and redo operations. Additionally, the research aims to advance the manipulation of multiple objects within images, using a single text prompt. For instance, a single prompt could instruct the model to add a bird to a branch and remove a rock, and these changes would be executed iteratively in a single cycle. This approach not only extends the scope of image editing based on textual prompts but also introduces efficiency through selective and simultaneous editing of multiple objects, thereby contributing to the advancement in the realm of image manipulation.

D.4: Details of Sub Objectives

The general objectives mentioned above in section D.3. can be broken down into the following sub-projects:

- Simultaneous Edits of Multiple Objects
- Undoing and Redoing Edits

Sub Project 1: Simultaneous Edits of Multiple Objects

D.4.1: Name of Researchers/Interns:

Table 1: Name of Researchers and Their Responsibilities

<table>
<tr><th>Name of Researcher</th><th>Education Level</th><th>Responsibilities</th><th>Expertise</th></tr>
<tr><td>Abhinav Chalise</td><td rowspan="4">Undergraduate</td><td rowspan="2">Efficient Implementation and Integration</td><td>Python, PyTorch, TensorFlow</td></tr>
<tr><td>Nimesh Gopal Pradhan</td><td>Python, TensorFlow, C++</td></tr>
<tr><td>Nishan Khanal</td><td>Performance Optimization and Robustness Testing</td><td>Python, PyTorch, Algorithm Optimization</td></tr>
<tr><td>Prashant Raj Bista</td><td>Mathematical Modeling For Simultaneous Edits</td><td>Linear Algebra, Probability, Natural Language Processing</td></tr>
<tr><td>Bishartha Manandhar</td><td>Graduate</td><td>Mathematical Modeling and Review of the theoretical portion</td><td>Mathematical modeling, Density functional theory-based simulation, Simulation in ARML lab</td></tr>
</table>

D.4.2: Specific objectives of the internship or subproject. Clearly state your [sub-] objectives so reviewers can assess if they are achievable.

Table 2: Sub-Objectives for Simultaneous Multi-Object Edits

Sub-Objective	Description
Extending Single Object Editing to Multiple Object Editing.	Extend the capabilities of existing single-object editing techniques to encompass multi-object editing.
Focusing on editing object color, spatial location of objects, insertion of foreign objects and removing existing objects.	Implement functionalities for editing multiple objects within an image, including adding, removing, changing object colors, and adjusting their spatial positions for enhanced visual customization.

Extending Single-Object Editing to Multi-Object

Extend the capabilities of traditional single-object editing approaches found in previous papers to the more complex realm of multi-object editing. This sub-objective seeks to build upon existing methodologies, allowing for a seamless transition from editing a single object to manipulating multiple objects concurrently.

Focusing on editing object color, spatial location of objects, insertion of foreign objects and removing existing objects

Develop functionalities that allow for the manipulation of multiple objects within an image. This research will encompass a wide range of edits, including the addition of foreign objects, removal of existing ones, and modification of color attributes for different objects. A key feature will be the ability to change the spatial locations of objects; for example, shifting a notebook from the right side to the left side of a table based on user prompts. The ultimate objective is to create a comprehensive and versatile multi-object editing system, responsive and adaptive to a variety of user directives.

D.4.3: Methodologies.

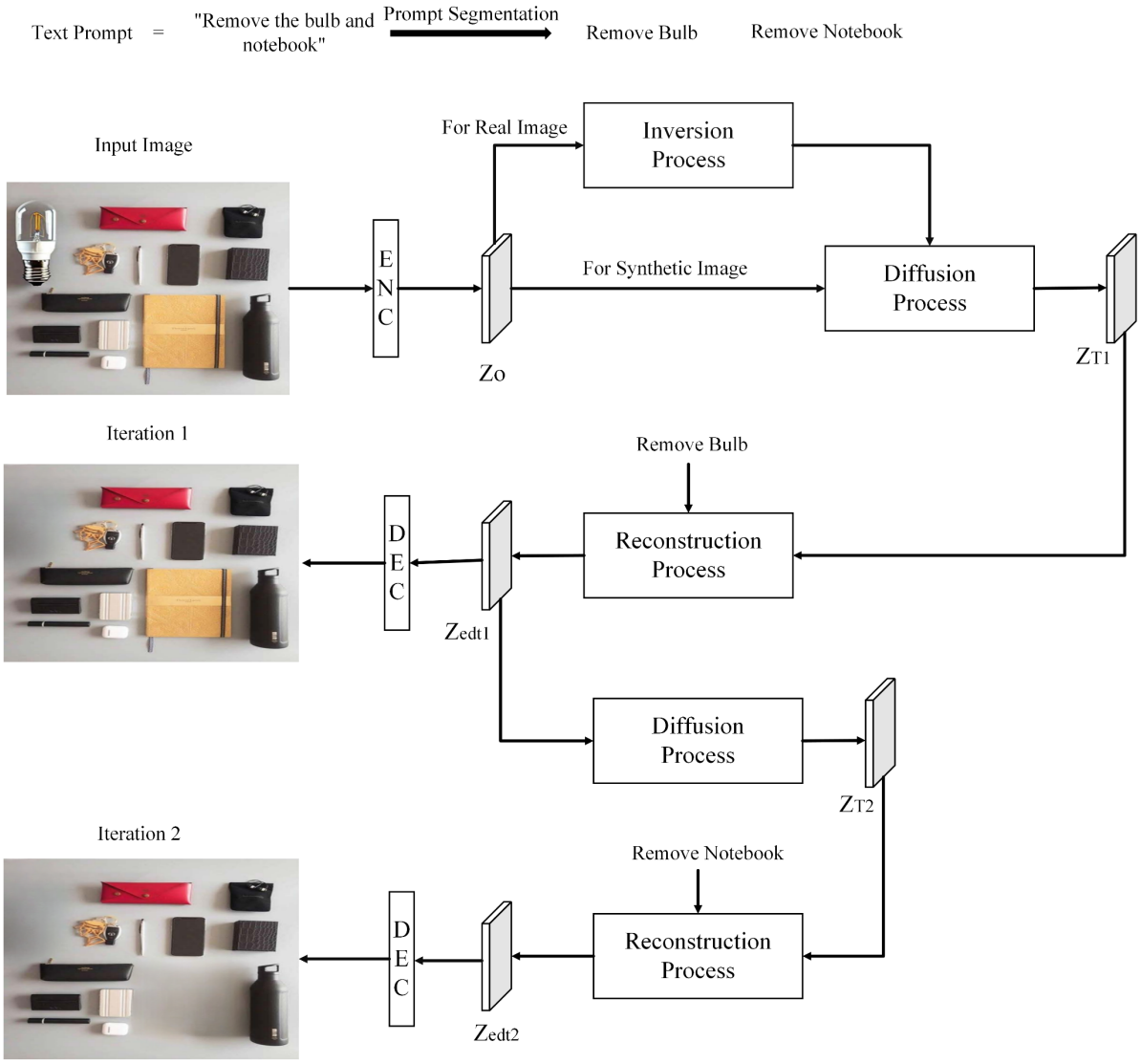


Figure 1: Conceptual Figure Showing Simultaneous Edits of Multiple Objects

Process Overview

An overview of our proposed image editing method. The first step is to encode the image to a latent space using an encoder where we get the first latent variable (Z_o) which is the input image represented in lower dimensions. For real images, Z_o is passed through the Inversion Process where the trajectory for reconstructing the given input image from noise is learned. In the case of a synthetic image, the trajectory is already known so it goes directly to the Diffusion channel where noise is gradually added to obtain the noisy latent variable (Z_T). The text prompt is also segmented into instructions. From the noisy latent the edited image is reconstructed by introducing the text embeddings to obtain the intermediate latent representation (Z_{edt}). This process is done iteratively until all the instructions given in the text prompt have been executed.

Model and Dataset Selection

For the research, selecting the right model and dataset is crucial for successful multi-object editing through text prompts. We opt for the pre-trained Stable Diffusion, a large-scale text-to-image model known for its proficiency in interpreting and manipulating images. This model, trained on extensive datasets, demonstrates excellent generalization capabilities. The text prompts are segmented and converted to text embeddings using the Contrastive Language-Image Pre-training (CLIP) in Stable Diffusion, aligning text prompts with the image editing process. For diverse scenarios and object types, we leverage freely available COCO and ImageNet datasets, ensuring the model's adaptability to various multi-object editing challenges.

Iterative Refinement and Editing

The Iterative Refinement in the proposed model involves a process where each editing step builds upon the previous one. Initially, the model applies the first set of changes based on the text prompt. It then stores the outcome which is used for further editing. This process repeats, with each iteration bringing the image closer to the desired outcome. This approach not only ensures that the final image accurately represents the prompt but also allows for enhancements or re-edits based on intermediate results, leading to a more precise and tailored output.

Instrumentation: Hardware and Software

For efficient execution, the model will rely on a powerful GPU like NVIDIA A100, over the CPU, enhancing computational speed for image editing tasks. GPU parallelism aligns with deep learning model architecture, accelerating both training and inference. For software, PyTorch and TensorFlow are chosen for community support, robust functionalities, and compatibility with GPU acceleration, ensuring efficient model implementation and overall performance optimization.

Evaluation Metrics

In evaluating our proposed method, we use metrics such as Learned Perceptual Image Patch Similarity (LPIPS) and the Multi-Scale Structured Similarity Indexing Method (MS-SSIM). These metrics excel in assessing perceptual and structural aspects more effectively than traditional measures. LPIPS measures perceptual similarity, aligning closely with human judgments, providing a more accurate evaluation of our model's edits in terms of human perception. MS-SSIM evaluates structural coherence, considering not just pixel-level details but also the arrangement and relationships between objects, providing insights into how well our model preserves the overall structure of the image.

Sub Project 2: Undoing and Redoing Edits

D.4.4: Name of Researchers/Interns:

Table 3: Name of Researchers and Their Responsibilities

Name of Researcher	Education Level	Responsibilities	Expertise
Abhinav Chalise	Undergraduate	Implementing in the program of Undo and Redo Functionalities	Python, PyTorch, TensorFlow
Nimesh Gopal Pradhan		Mathematical Modeling for Editing Control	Linear Algebra, Calculus
Prashant Raj Bista			Linear Algebra, Probability
Nishan Khanal		Implementation of Iterative Feedback System	Python, Pytorch
Bishartha Manandhar	Graduate	Mathematical Modeling and Review of the theoretical portion	Mathematical modeling, Density functional theory-based simulation, Simulation in ARML lab

D.4.5: Specific objectives of the internship or subproject. Clearly state your [sub-] objectives so reviewers can assess if they are achievable.

Table 4: Sub-Objectives for Undoing and Redoing Edits

Sub-objective	Description
Design of State Management System	Create an efficient system for managing latent representations during image editing iterations, ensuring seamless storage and retrieval.
Removal of artifacts when performing undo/redo operation	Create mechanisms to remove artifacts during undo/redo, ensuring a clean and accurate image restoration.
Reversing and redoing up to two edits	Enables users to undo and redo two previous editing operations, enhancing flexibility and user-friendly editing.

Design of State Management System

Creating an efficient system to manage latent representations during the editing. This involves storing latent representation after each iteration and implementing a system that can intelligently locate the necessary latent representation/s based on the provided text prompt for facilitating precise navigation for users.

Removal of Artifacts in Undo/Redo Operations

Refining the undo/redo operations by addressing and eliminating potential artifacts such as image distortion, blurred objects, and instances where undoing fails which are issues identified in previous research. The ultimate goal is to guarantee that undoing or redoing image edits is a smooth and distortion-free process, for a visually satisfying editing experience.

Reversing and Redoing up to Two Edits

Incorporating functionality that allows users to undo and redo at least two previous editing operations. Users should be able to undo edits for multiple objects in a non-contiguous manner, while redo operations focus on a single object at a time in a contiguous fashion, enhancing control over the editing history.

D.4.6: Methodologies.

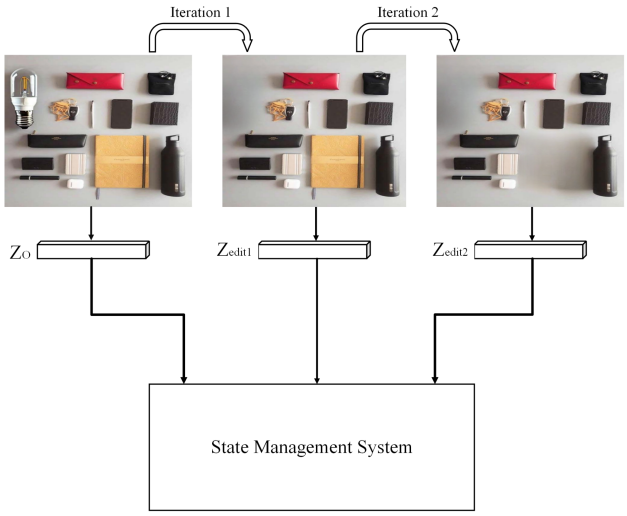


Figure 2: Conceptual Figure Showing State Management System

State Management System

The State Management System plays the main role in preserving the iterative history of image edits. As the image undergoes transformations in various iterations, the system captures and retains the latent representation at each stage. For example, in Iteration 1, the removal of the bulb and, in Iteration 2, the removal of the notebook generate distinct latent representations. These representations are stored by the state management system, forming an archive of the image's evolution. This repository of latent representations is used for undo and redo operations.

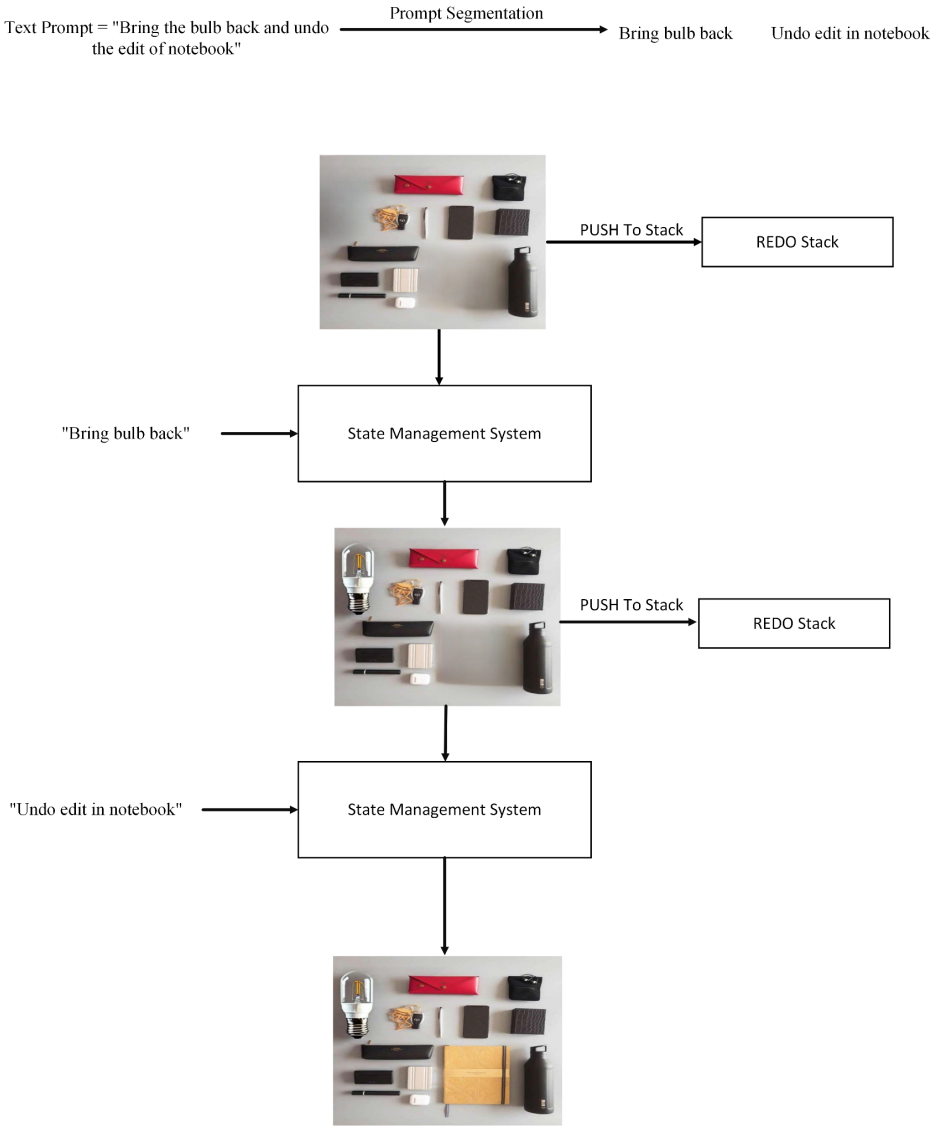


Figure 3: Conceptual Figure Showing Undo Functionality

Undo Functionality

When a user issues an Undo prompt, the state management intelligently locates the most appropriate latent representation/s from the stored iterative stages based on the text prompt. Subsequent changes are then applied to these selected latent representation/s, and through the decoding process, the desired Undo effect is achieved. The current latent representation is stored in a separate stack known as the Redo stack. The Undo operation isn't confined to a strict chronological order. Users have the flexibility to selectively retain changes for certain objects while undoing alterations for others. For instance, in the context of our iterative stages, a user could choose to undo edits for the bulb (Iteration 1) while retaining the edit for the notebook (Iteration 2). This approach allows users to tailor their editing history according to specific objects of the image. The text prompt of the Undo mechanism can also affect multiple objects at the same time.

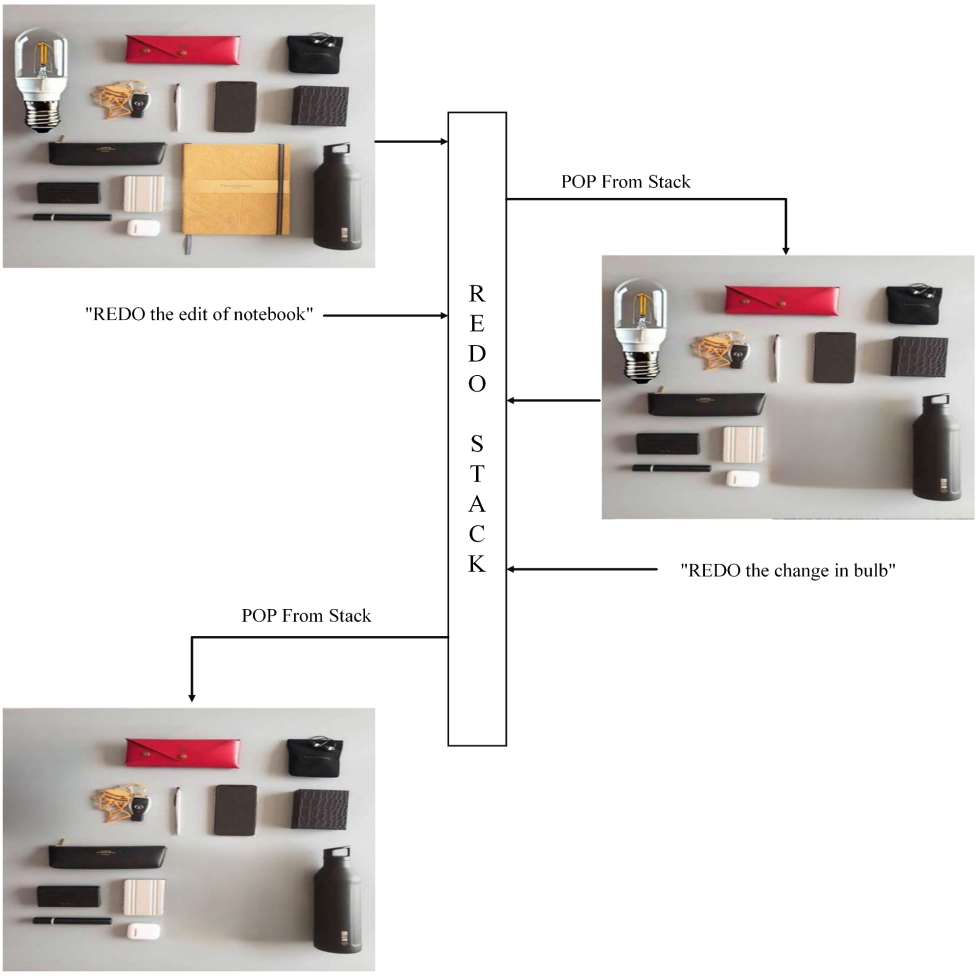


Figure 4: Conceptual Figure Showing Redo Functionality

Redo Functionality

When a user initiates a Redo prompt, the State Management System retrieves the latent representation stored in the Redo stack. This latent representation, previously saved during the Undo operation, is then popped from the stack and decoded. The result is a precise replication of the desired edit, effectively redoing the user's selected changes. The Redo mechanism allows users to revisit and reapply edits that were previously undone. This iterative approach empowers users to navigate through their editing history, ensuring that no creative decisions are lost. The redo operation will be done in a contiguous fashion, unlike the undo operation.

Evaluation Metric

The evaluation of the UNDO and REDO mechanisms in our system adopts a dual approach, combining both qualitative and quantitative assessments. Qualitatively, it heavily relies on user feedback to gauge the perceptual quality of images after UNDO and REDO operations are applied. This user-centric approach is crucial for understanding the visual integrity and detecting any artifacts that might arise from these operations. On the quantitative side, the evaluation ensures that the system supports a minimum of two levels of undo functionality. Additionally, we use CLIP scores as a metric to quantitatively measure how well the system understands and executes text prompts, assessing the alignment between the given instructions and the resultant image edits. The evaluation process also includes checking the system's ability to selectively undo changes for individual objects without affecting other alterations in the image, ensuring a nuanced and effective UNDO mechanism.

D.5: Timeline including Gantt chart

Activity 1. Pre-screening of Proposal

Introduction to the research idea, showing past research, and presenting novel ideas for the development of the research landscape.

Activity 2. Detailed Proposal Submission

Extension of the pre-screened proposal to demonstrate the uniqueness of the research problem, emphasizing the gap in previous solutions and stating the need for the proposed research.

Activity 3. Simultaneous Multi-object Editing

Implementing and testing simultaneous image editing involving the formulation of advanced algorithms and mathematical models for the diffusion model.

Activity 4. Undo and Redo of Edits

Implementing and testing undo and redo functionalities, driven by the development of precise algorithms and mathematical aspects to enhance the model’s flexibility.

Activity 5. Web Interface

Developing a webpage interface ensuring user-friendly accessibility for a diverse audience, enhancing the model's usability and practicality

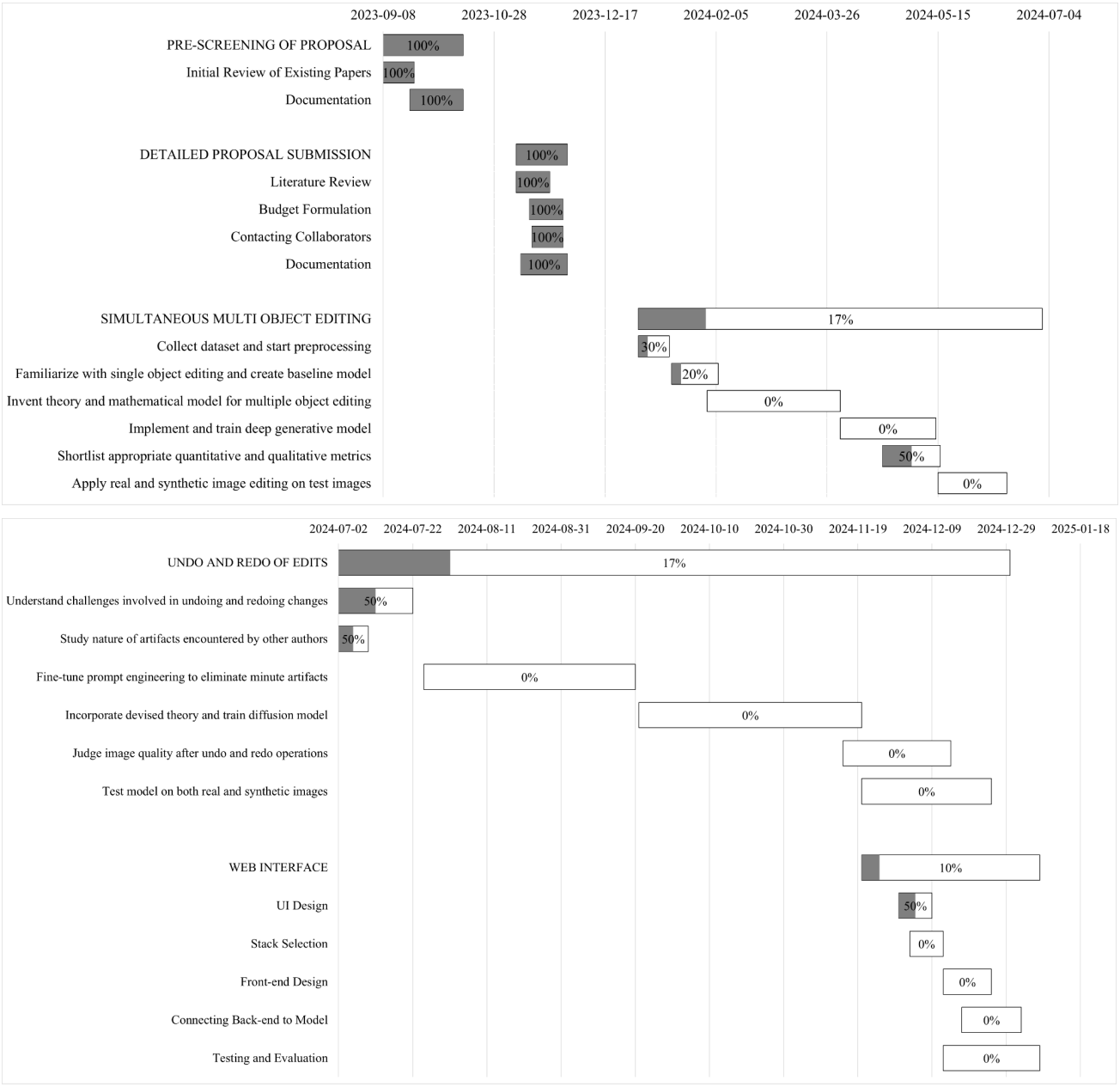


Figure 5: Gantt Chart

D.6: Milestone Table for Work Plan and Deliverables

Table 5: Work Plan and Deliverables

Activity	Risk(s): Description, probability, and impact	Risk mitigation strategies
Finding and Learning Existing Diffusion Models	Risk: Proprietary and paid models may hinder from evaluation of diverse models Probability: Low Impact: Difficulty in comparing results with different models	1. Most models like Stable Diffusion are open-source and can be obtained from GitHub projects, Hugging Face, and Kaggle. 2. Proprietary resources that may be required can be bought.
Programming and Training Models to Implement Multi-Object Editing	Risk: Lack of computation Power to train and test models in the relevant time Probability: Medium Impact: Failure in developing the proposed solution	1. Applying for credits and free resources provided by cloud computing solutions. 2. Cloud computing solutions can be obtained from Azure or AWS if required.
Developing Mathematical Models For Simultaneous Edits and Undo/Redo Functionality	Risk: Incorrect mathematical modeling or failure to develop the intended models. Probability: Medium Impact: Delay in research completion	1. Seek guidance from previous research publishers, and conduct regular peer reviews to validate and refine the mathematical models

D.7: Budget Plan

D.7.1: Personnel salaries and stipends

In this research, all 6 participants will be provided with a stipend of Rs 9,100 each, which will be around 27.5% of the total budget provided by the grant. A 2.5% of the total budget will be provided as a stipend to fellow participants who helped during the research. So, 30% of the total budget is allocated as a stipend.

D.7.2: Research materials and equipment

For this research, materials and equipment will cost about Rs. 60,000. Our main uses will be on GPU for training our model, which will be 30% of the total budget. The cost of the NVIDIA TESLA V100 GPU on Azure is around $6.12/hour. The cost here will be utilized if we require any software or any other paid service for the completion of the task. In this research, huge storage will be required, for the storage and RAM budget will be utilized from here.

D.7.3: Travel and accommodation for conferences and collaborations

This research work will be shown at various graduate conferences. Some research graduate conferences include the graduate conferences of Pulchowk Campus, Kathmandu University, and WRC. For this purpose, 20% of the total budget is allocated i.e. Rs.40,000.

D.7.4: Data collection and analysis and publication costs

The data this research uses is free of cost. This research work will be published in various journals local and international. Publications costs include 10% of the total budget i.e. Rs.20,000.

D.7.5: Other justifiable research expenses

For other miscellaneous costs, we have allocated 10% of the total budget.

Table 6: Budget Allocation Plan

Budget Plan	Allocated Amount
Personnel salaries and stipends	Rs. 60,000
Research materials and equipment	Rs. 60,000
Travel and accommodation for conferences and collaborations	Rs. 40,000
Data collection and analysis and publication costs	Rs. 20,000
Other justifiable research expenses	Rs. 20,000
Total	**Rs. 2,00,000**

D.8: Expected Final Deliverables (Final Research Output and its Potential Impact, Journal and Conference Publications, and similar)

Web Interface Creation

Upon completion, we will provide an intuitive web interface for providing the input image and edit prompt. This page will generate the requested edit within some time frame and provide an option to download the result. The images below show a possible web interface for the model which is easy to use and requires no technical knowledge of the underlying model.

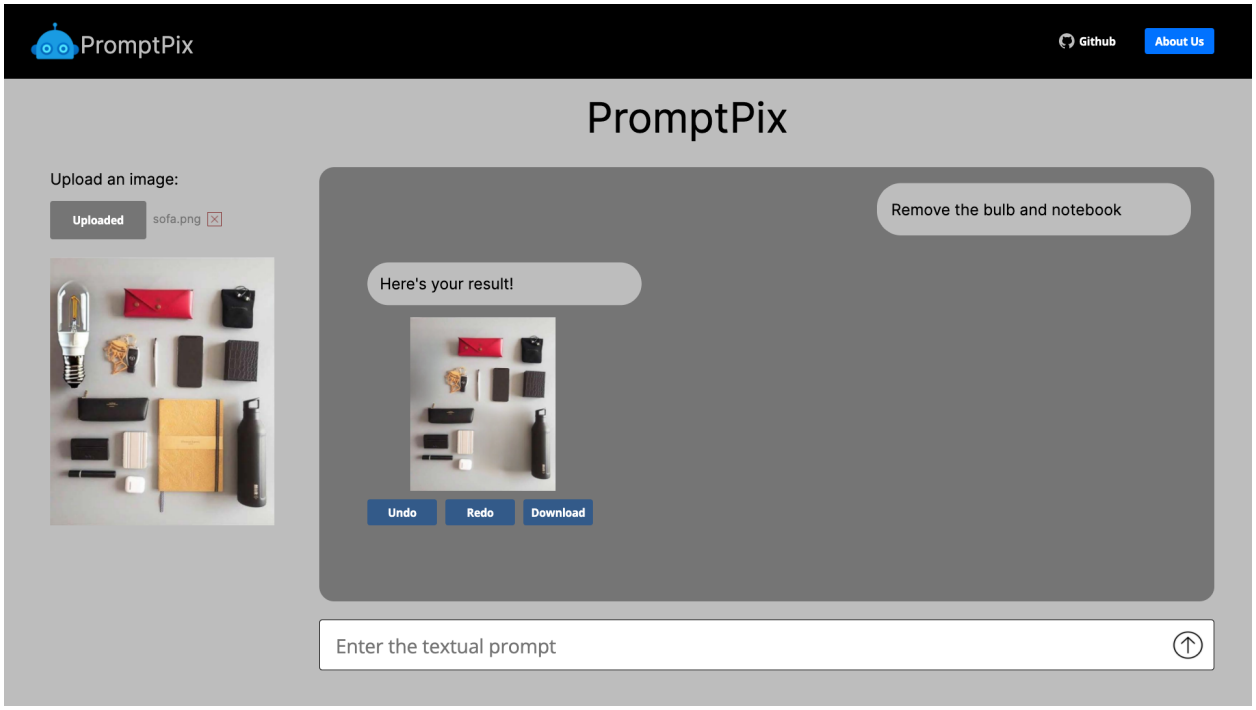


Figure 6: A conceptual web interface to upload a local image and write prompts to edit the image

Impacts to Society

The integration of a deep generative model for text-guided image manipulation has the potential to revolutionize how visual content is conceptualized, created, and communicated across various industries, fostering efficiency, creativity, and inclusivity in the process.

Animation studios like Incessant Rain can have streamlined early stages of production by allowing animators to input text descriptions or instructions to generate initial visuals and designs. This can rapidly speed up the conceptualization process. Ideas and iterative changes can be communicated easily through textual descriptions allowing for quicker iterations and consensus among the team. Morphing from one scenario to another can be done in very unique and creative ways through the use of generative models.

Likewise, content creators and graphics designers are also able to create their desired graphics like creating thumbnails for their YouTube videos through textual descriptions and base image context. Textual narratives can also be converted to visual elements, enriching the storytelling.

All in all, reducing the need for multiple iterations and cutting down production timelines and costs, allowing a broader range of creators to make these fields more accessible, and also helping educators through applications in educational materials have the potential to positively impact society.

Journal Section

We will initially submit our paper to the NCE Journal of Science & Engineering (NJSE), a national journal that fosters diverse research content. Established in 2020 by the National College of Engineering, NJSE will serve as a solid foundation for our research to gain national recognition.

After publication in NJSE, we plan to further extend our research, exploring new processes and ideas. This will prepare us for submission to prominent international journals. Our targets include the International Journal of Pattern Recognition and Artificial Intelligence, the International Journal of Artificial Intelligence, and Procedia Computer Science. Although these journals need publication costs, we plan to publish them with open access. These journals, with their global reach and varied focus on AI and Computer Science, will be ideal for the expanded scope and depth of our evolved research. Procedia Computer Science, in particular, offers an appealing platform due to its free publication policy and open access availability, complemented by a high CiteScore of 4. This phased approach from national to international publication will ensure a comprehensive dissemination and impactful presentation of our work.

D.9: Research Basis and Working Conditions to Undertake the Proposed Research

D.9.1: Research Basis

This project sits at the intersection of **computer vision** and **natural language processing** and hinges on prompt engineering, an approach to adapting a large pre-trained model to new tasks by augmenting the model input with task-specific hints. Specifically, this project focuses on image editing using texts to achieve intuitive and versatile modification of images. As the principal investigator of this project, I have supervised both bachelor's degree projects and master's degree theses that pertain to domains within the fields of computer vision and natural language processing. Some of the research works that I supervised have resulted in articles that have been published in peer-reviewed journals.

(1) The article entitled "***Soccer Game Summarization using Audio Commentary, Metadata, and Captions***" was published in the ***Proceedings of the 1st Workshop on User-centric Narrative Summarization of Long Videos*** in 2022. The focus of this research work was on the generation of complete soccer game summaries in continuous text format with length constraints, based on raw game multimedia, as well as readily available game metadata and captions, using **Natural Language Processing** along with heuristics.

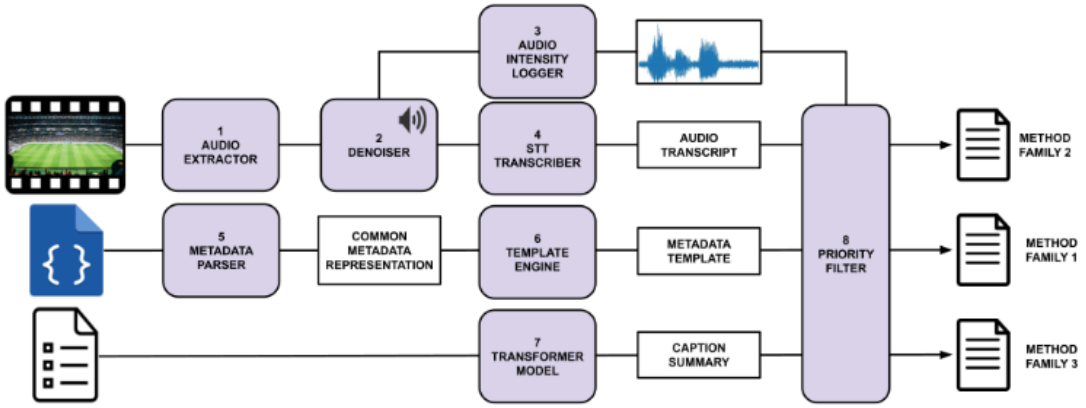


Figure 7: Pipeline for Generating Text Summaries from Soccer Game Multimedia

Table 7: Template for Generating Naïve Interpretations from Metadata

Dataset	Metadata Format	Interpretation	Sample Sentence
HOST	('free_kick', 'offending_player', 'team')	d[team][value] was awarded a free kick because of d[offending_player][value].	Bodø/Glimt was awarded a free kick because of Erling Haaland.
SoccerNet	('free_kick', 'team')	d[team][value] was awarded a free kick.	Bodø/Glimt was awarded a free kick.
HOST	('red_card', 'player', 'team')	d[player][value] from d[team][value] got a red card.	Sondre Sørli from Bodø/Glimt got a red card.
SoccerNet	('red_card', 'team')	d[player][value] got a red card.	Bodø/Glimt got a red card.
HOST	('goal', 'assist_by', 'scorer', 'shot_type', 'team')	d[scorer][value] scored a goal by d[shot_type][value] shot for d[team][value] with assistance from d[assist_by][value].	Sondre Sørli scored a goal by right-footed shot for Bodø/Glimt with assistance from Japhet Sery.
SoccerNet	('goal', 'team')	d[team][value] scored a goal.	Bodø/Glimt scored a goal.
HOST	('substitution', 'player_in', 'player_out', 'team)	d[player_in][value] replaced d[player_out][value] in d[team][value].	Patrick Berg replaced Sondre Sørli in Bodø/Glimt.
SoccerNet	('substitution', 'team)	d[player_in][value] replaced one of its players.	Bodø/Glimt replaced one of its players.

Figure 8: Audio Analysis Dashboard Showing Correlation of Audio Intensity Levels and Game Events

(2) The article entitled "***Assisting Soccer Game Summarization via Audio Intensity Analysis of Game Highlights***" was published in the ***Proceedings of the 12th IOE Graduate Conference*** in 2022. This research work utilized **signal processing** and **natural language processing** to explore the properties of audio signals during soccer events to calculate excitement among the audience. Enabling the calculation of excitement around different soccer events helped filter game highlights, and supported automatic video summarization based on the audience's perceived importance.

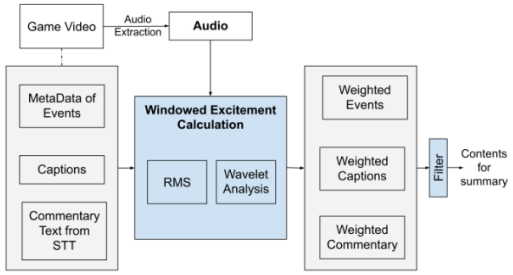


Figure 9: Audio-based Filter in Summarization Pipeline

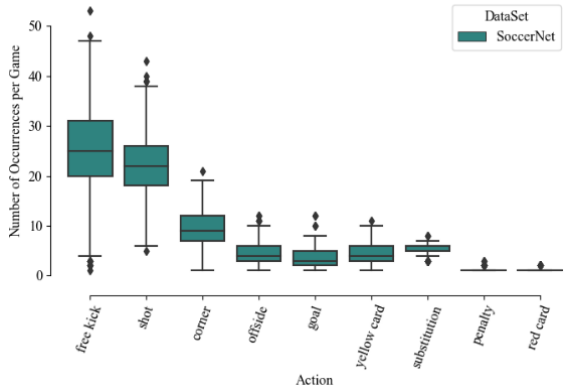


Figure 10: Distribution of Different Types of Events per Game in SoccerNet Dataset

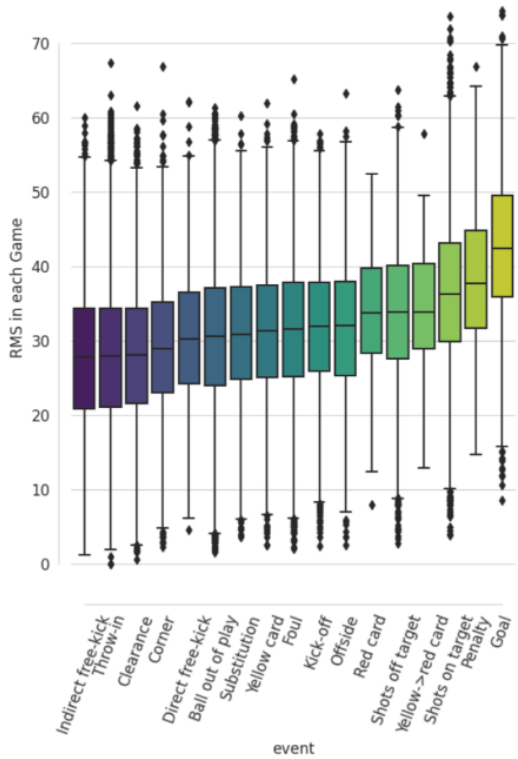


Figure 11: Audio RMS Distribution of Different Soccer Events

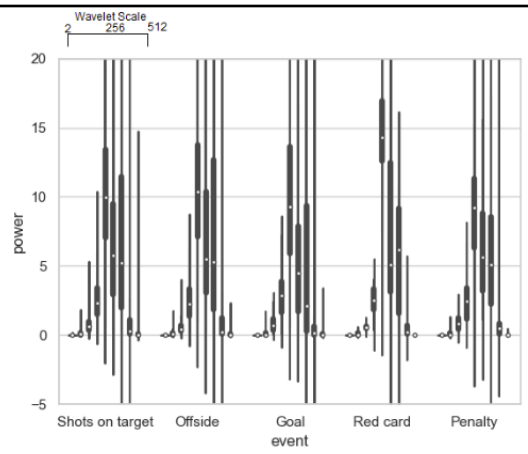


Figure 12: Power distribution across Wavelet scales

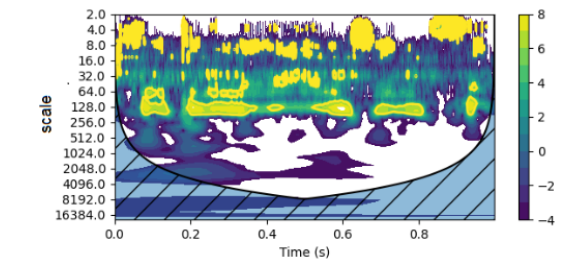


Figure 13: Wavelet Power Spectrum

(3) The article entitled "***Silent Speech Recognition in Nepali***" was published in the ***Proceedings of the 12th IOE Graduate Conference*** in 2022. This research work makes use of **computer vision** and **signal processing** techniques to provide a secure and seamless interaction between a human and a computer using silent speech recognition via surface electromyography (sEMG) signals recorded from the facial muscles of a speaker. The spectrogram of processed sEMG signals obtained from 8-channel gold cup electrodes is used to train a Convolution Neural Network (CNN). The trained model is then deployed to predict the silent utterances.

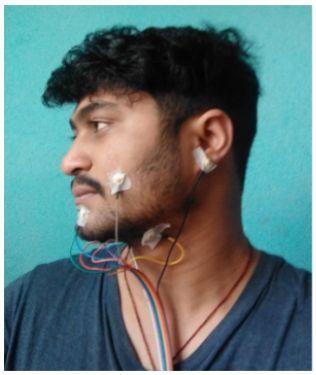


Figure 14: Electrode Placement

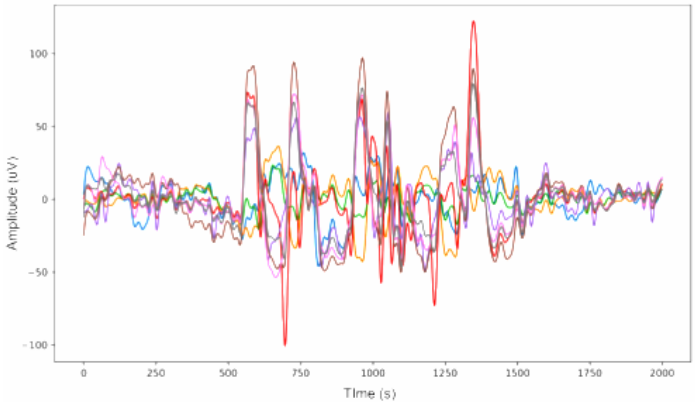


Figure 15: Time Domain Waveform of Filtered sEMG Signal Obtained from 8-channel Electrodes for the Utterance "अबको समय सुनाउ"

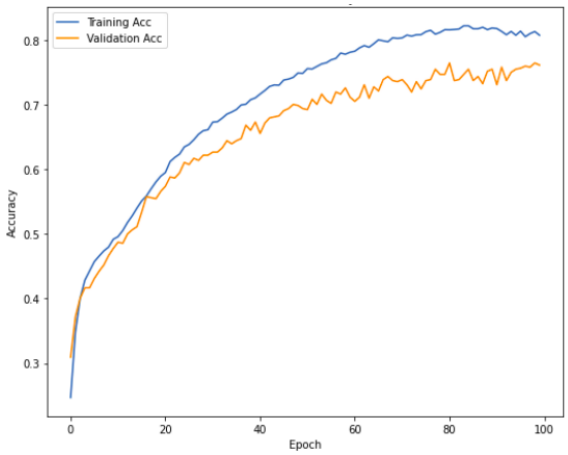


Figure 16: Performance of CNN Model

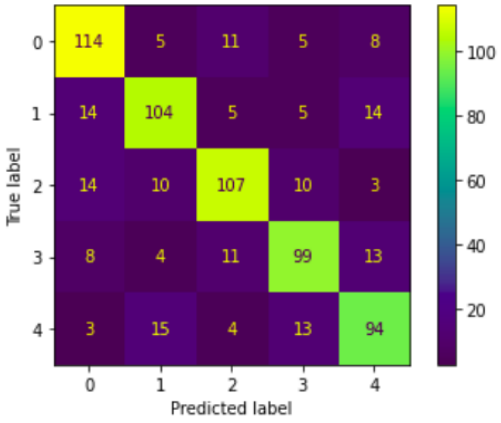


Figure 17: Confusion Matrix of the Model. The Labels 0, 1, 2, 3, and 4 respectively denote the Sentences "बत्तिको अवस्था बदल" "आजको मौसम बताउ" "एउटा संङ्गित बजाउ" "पङ्खाको स्थिति बदल" and "अबको समय बताउ"

D.9.2: Working Conditions

Thapathali Campus, IOE, TU is equipped with computer laboratories capable of handling basic tasks and testing small-scale algorithms. However, these computers lack the necessary computing power for intensive tasks such as large-scale modeling, simulations, and data training.

Figure 18: Computer Lab at Thapathali Campus, IOE, TU

Table 8: Specifications of Computers Lab at Thapathali Campus, IOE, TU

Processor	Memory	Storage	Cores	Threads	Graphics	Operating System
9th Generation Intel Core i5 Processor	8 GB	1 TB	4 cores	8 threads	Intel UHD Graphics 630	Windows 10 Pro
9th Generation Intel Core i3 Processor	4 GB	500 GB	4 cores	4 threads	Intel UHD Graphics 630	Windows 10 Pro

To address the limitations in experimental conditions, the researchers have initiated a collaborative effort with the Advanced Materials Research Laboratory at CDP. This partnership is a significant step forward, as the laboratory has agreed to furnish the necessary computational resources. Notably, they provide access to the NVIDIA A100 Tensor Core GPU, ensuring the smooth progression of our experimental work. Below is the letter of intent from the Advanced Materials Research Laboratory at CDP, which serves as a formal testament to our collaborative engagement.

Figure 19: Workstation at AMRL, CDP, TU

Figure 20: Stacks of servers at AMRL, CDP

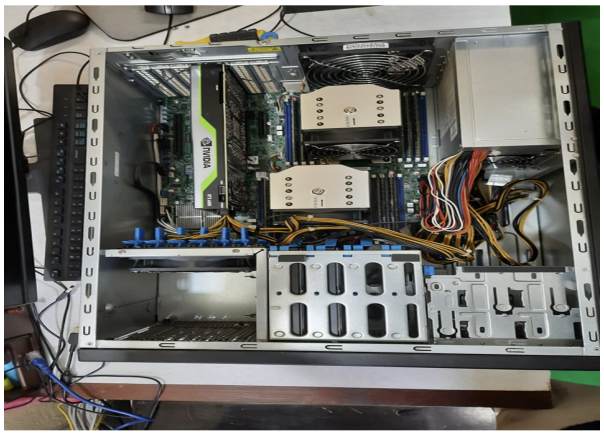
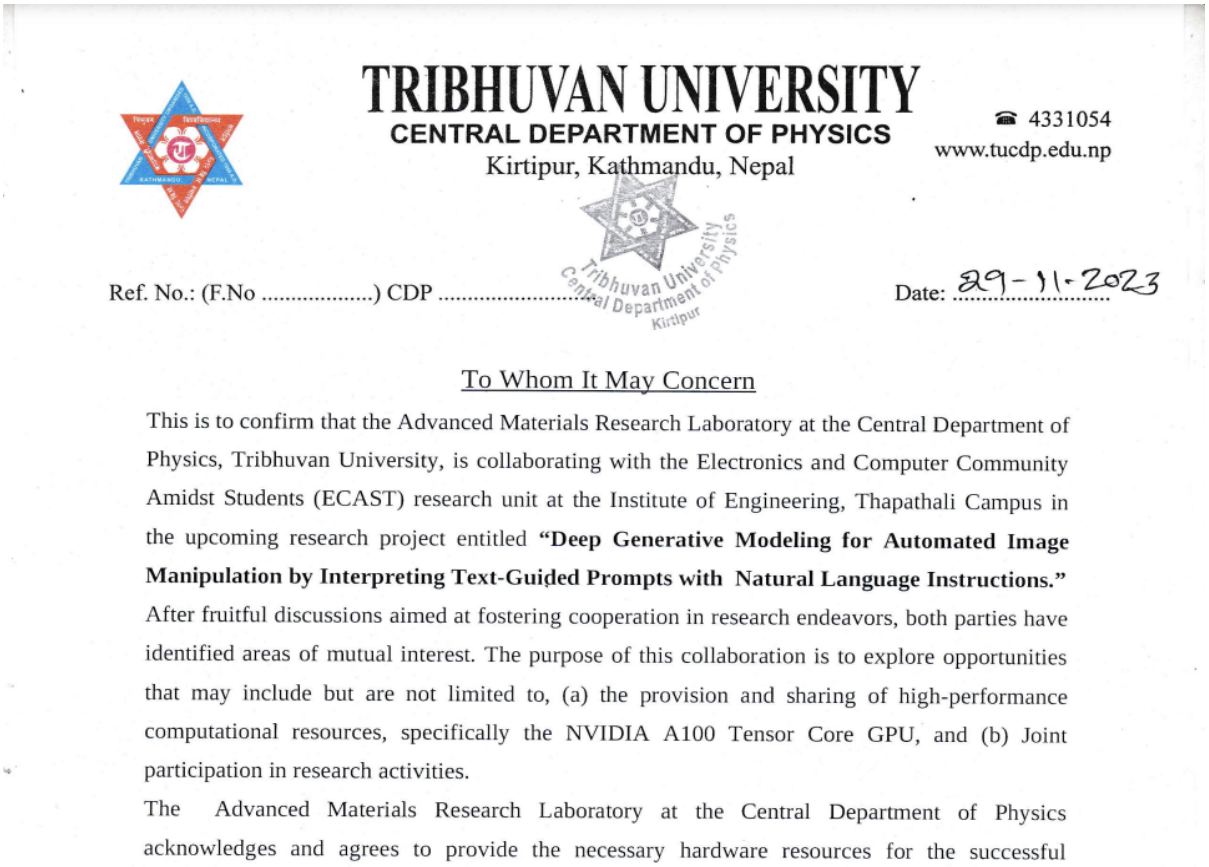


Figure 21: Workstation with NVIDIA RTX 6000 at AMRL, CDP

Table 9: Specifications of High-End Computing Machines at AMRL, CDP

GPU	GPU Memory	CUDA Core count	Tensor core count	RAM
NVIDIA A100 Tensor Core GPU	160 GB	27,648	1,728	DDR4, 512 GB
NVIDIA RTX 6000	24 GB	4,608	576	DDR6, 24 GB



TRIBHUVAN UNIVERSITY

CENTRAL DEPARTMENT OF PHYSICS

Kirtipur, Kathmandu, Nepal

☎ 4331054
www.tucdp.edu.np

Ref. No.: (F.No .....) CDP .....

Tribhuvan University Central Department of Physics Kirtipur

Date: 29-11-2023

To Whom It May Concern

This is to confirm that the Advanced Materials Research Laboratory at the Central Department of Physics, Tribhuvan University, is collaborating with the Electronics and Computer Community Amidst Students (ECAST) research unit at the Institute of Engineering, Thapathali Campus in the upcoming research project entitled **"Deep Generative Modeling for Automated Image Manipulation by Interpreting Text-Guided Prompts with Natural Language Instructions."** After fruitful discussions aimed at fostering cooperation in research endeavors, both parties have identified areas of mutual interest. The purpose of this collaboration is to explore opportunities that may include but are not limited to, (a) the provision and sharing of high-performance computational resources, specifically the NVIDIA A100 Tensor Core GPU, and (b) Joint participation in research activities.

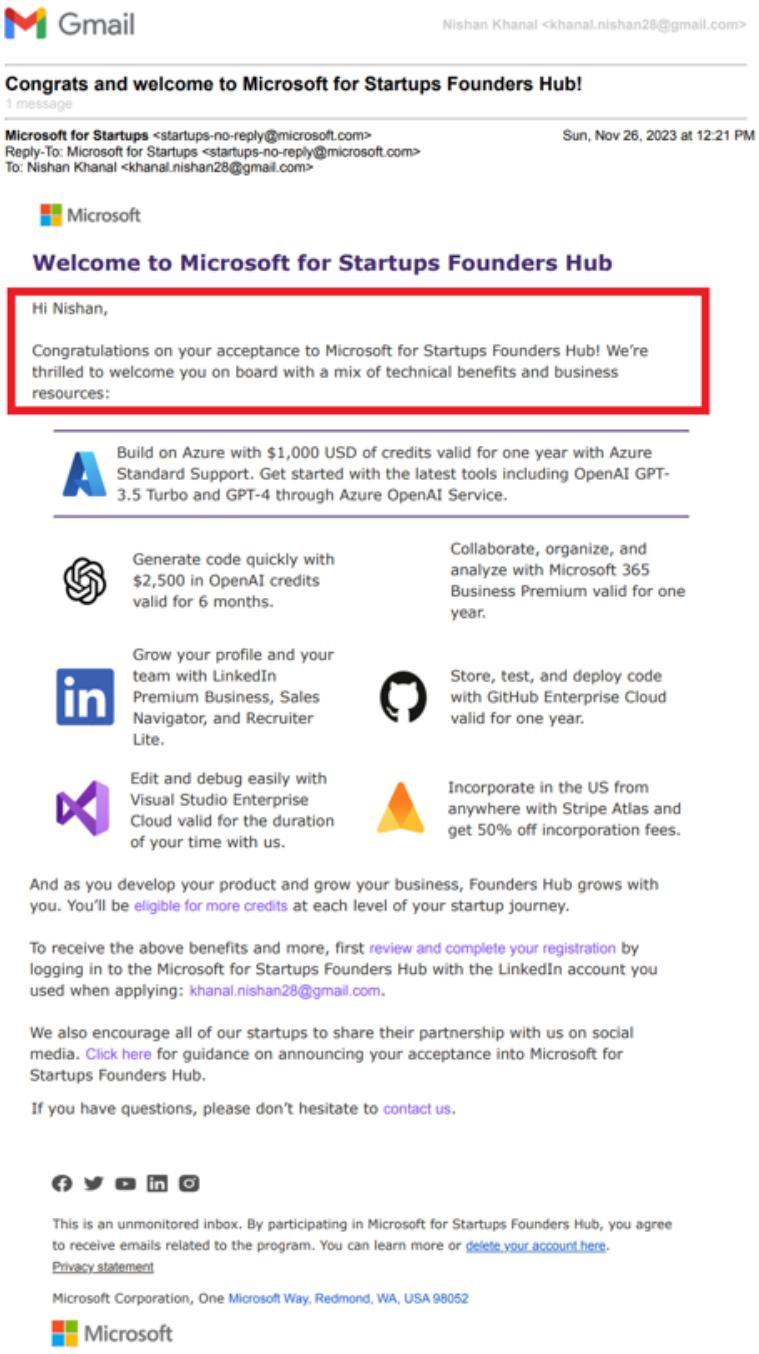
The Advanced Materials Research Laboratory at the Central Department of Physics acknowledges and agrees to provide the necessary hardware resources for the successful execution of the collaborative research initiatives.

This letter of intent is valid for the duration from 01/01/2024 to 14/07/2025 and is subject to the approval and endorsement of the authorities within the Central Department of Physics.

Madhav Prasad Ghimire, PhD
Associate professor &
Head
Advanced Materials Research Laboratory
email: madhav.ghimire@cdp.tu.edu.np

Figure 22: Letter Of Intent to Collaborate From Advanced Materials Research Laboratory at the Central Department Of Physics, Tribhuvan University

The research team has successfully applied for resources through Microsoft's Startups Founder Hub program, securing $1,000 in Azure credits. Attached below is the official confirmation from Microsoft, evidencing the receipt of these credits. This generous allocation of $1000 in Azure credits will be strategically utilized to enhance our cloud computing capabilities, further bolstering the efficiency and scope of our research endeavors.



Gmail — Nishan Khanal <khanal.nishan28@gmail.com>

Congrats and welcome to Microsoft for Startups Founders Hub!

1 message

Microsoft for Startups <startups-no-reply@microsoft.com> Sun, Nov 26, 2023 at 12:21 PM
Reply-To: Microsoft for Startups <startups-no-reply@microsoft.com>
To: Nishan Khanal <khanal.nishan28@gmail.com>

Microsoft

Welcome to Microsoft for Startups Founders Hub

Hi Nishan,

Congratulations on your acceptance to Microsoft for Startups Founders Hub! We're thrilled to welcome you on board with a mix of technical benefits and business resources:

Build on Azure with $1,000 USD of credits valid for one year with Azure Standard Support. Get started with the latest tools including OpenAI GPT-3.5 Turbo and GPT-4 through Azure OpenAI Service.

Generate code quickly with $2,500 in OpenAI credits valid for 6 months.

Collaborate, organize, and analyze with Microsoft 365 Business Premium valid for one year.

Grow your profile and your team with LinkedIn Premium Business, Sales Navigator, and Recruiter Lite.

Store, test, and deploy code with GitHub Enterprise Cloud valid for one year.

Edit and debug easily with Visual Studio Enterprise Cloud valid for the duration of your time with us.

Incorporate in the US from anywhere with Stripe Atlas and get 50% off incorporation fees.

And as you develop your product and grow your business, Founders Hub grows with you. You'll be eligible for more credits at each level of your startup journey.

To receive the above benefits and more, first review and complete your registration by logging in to the Microsoft for Startups Founders Hub with the LinkedIn account you used when applying: khanal.nishan28@gmail.com.

We also encourage all of our startups to share their partnership with us on social media. Click here for guidance on announcing your acceptance into Microsoft for Startups Founders Hub.

If you have questions, please don't hesitate to contact us.

This is an unmonitored inbox. By participating in Microsoft for Startups Founders Hub, you agree to receive emails related to the program. You can learn more or delete your account here.
Privacy statement

Microsoft Corporation, One Microsoft Way, Redmond, WA, USA 98052

Microsoft

Figure 23: Mail from Microsoft Confirming $1,000 Azure Credit

The $1,000 worth of Azure credit can be used for cloud computing as follows:

Table 10: GPU with their Price per Hour

GPU	vCPUs	RAM	Price per hour ($)	Hours of use
NVIDIA Tesla T4	16	110 GiB	1.204/hour	830
NVIDIA Tesla V100	6	112 GiB	3.06/hour	326
NVIDIA Tesla V100	12	224 GiB	6.12/hour	163

D.10: Benefit to the Faculties and Undergraduate Students

D.10.1: Benefit to the faculties

This research can help people across different faculties. For civil and architectural engineering, the model can help in visualizing design concepts and planning. Students can describe their design ideas through text prompts, generating visual representations that can facilitate communication and collaboration. In industrial engineering, students can prototype their ideas to generate visual prototypes and simulate changes in manufacturing processes. In summary, this research will provide a model that can serve as a versatile tool across different engineering disciplines, offering benefits in design, visualization, communication, and collaborative exploration of creative ideas.

D.10.2: Benefit to the undergraduate students

This model can serve as an educational tool for students studying computer vision, deep learning, and image processing. It provides a practical example of how advanced technologies can be applied to creative tasks. Undergraduate students can use this model for the hands-on learning experience, by experimenting with different text prompts to see how the model reacts. Students can also test their new ideas with the model. By providing a text-based interface for image editing, the model makes image editing more accessible to individuals who are not familiar with design software. This model will help the undergraduates to make edits faster and easier.

Appendix-1: Proof of Affiliation for Faculty.

त्रिभुवन विश्वविद्यालय
Tribhuvan University
इन्जिनियरिङ्ग अध्ययन संस्थान
Institute of Engineering
थापाथली क्याम्पस
THAPATHALI CAMPUS

GPO Box-280, Thapathali, Kathmandu
Tel: 4-246465, 4218300, Fax: 977-1-4247340
E-mail: info@tcioe.edu.np
Website: www.tcioe.edu.np
गोश्वारा पो.ब.नं. २८०, थापाथली, काठमाडौं
फोन-४२४६४६५, ४२४६३०७
फ्याक्स: ५७७-१-४२४७३४०

Date:- 28/11/2023

To Whom It May Concern

This is to confirm that Mr. Dinesh Baniya Kshatri is currently affiliated with the Institute of Engineering, Thapathali Campus in Kathmandu, Nepal.

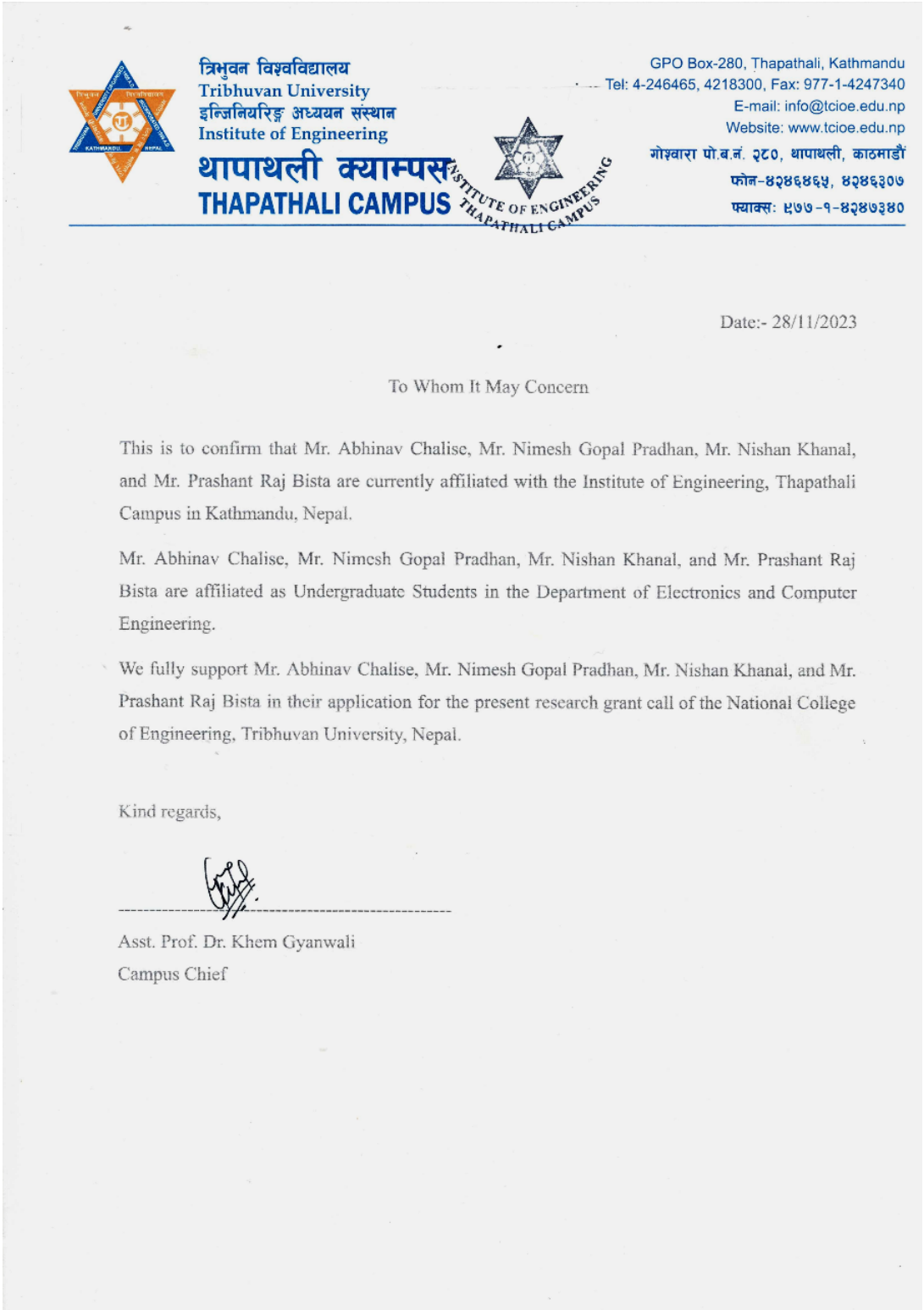
Mr. Dinesh Baniya Kshatri is affiliated as an Assistant Professor in the Department of Electronics and Computer Engineering.

We fully support Mr. Kshatri in his application for the present research grant call of the National College of Engineering, Tribhuvan University, Nepal.

Kind regards,

Asst. Prof. Dr. Khem Gyanwali
Campus Chief

Appendix-2: Proof of Affiliation for Undergraduate Students.



त्रिभुवन विश्वविद्यालय
Tribhuvan University
इन्जिनियरिङ्ग अध्ययन संस्थान
Institute of Engineering
थापाथली क्याम्पस
THAPATHALI CAMPUS

GPO Box-280, Thapathali, Kathmandu
Tel: 4-246465, 4218300, Fax: 977-1-4247340
E-mail: info@tcioe.edu.np
Website: www.tcioe.edu.np
गोश्वारा पो.ब.नं. २८०, थापाथली, काठमाडौं
फोन-४२४६४६५, ४२४६३०७
फ्याक्स: ५७७-१-४२४७३४०

INSTITUTE OF ENGINEERING THAPATHALI CAMPUS

Date:- 28/11/2023

To Whom It May Concern

This is to confirm that Mr. Abhinav Chalise, Mr. Nimesh Gopal Pradhan, Mr. Nishan Khanal, and Mr. Prashant Raj Bista are currently affiliated with the Institute of Engineering, Thapathali Campus in Kathmandu, Nepal.

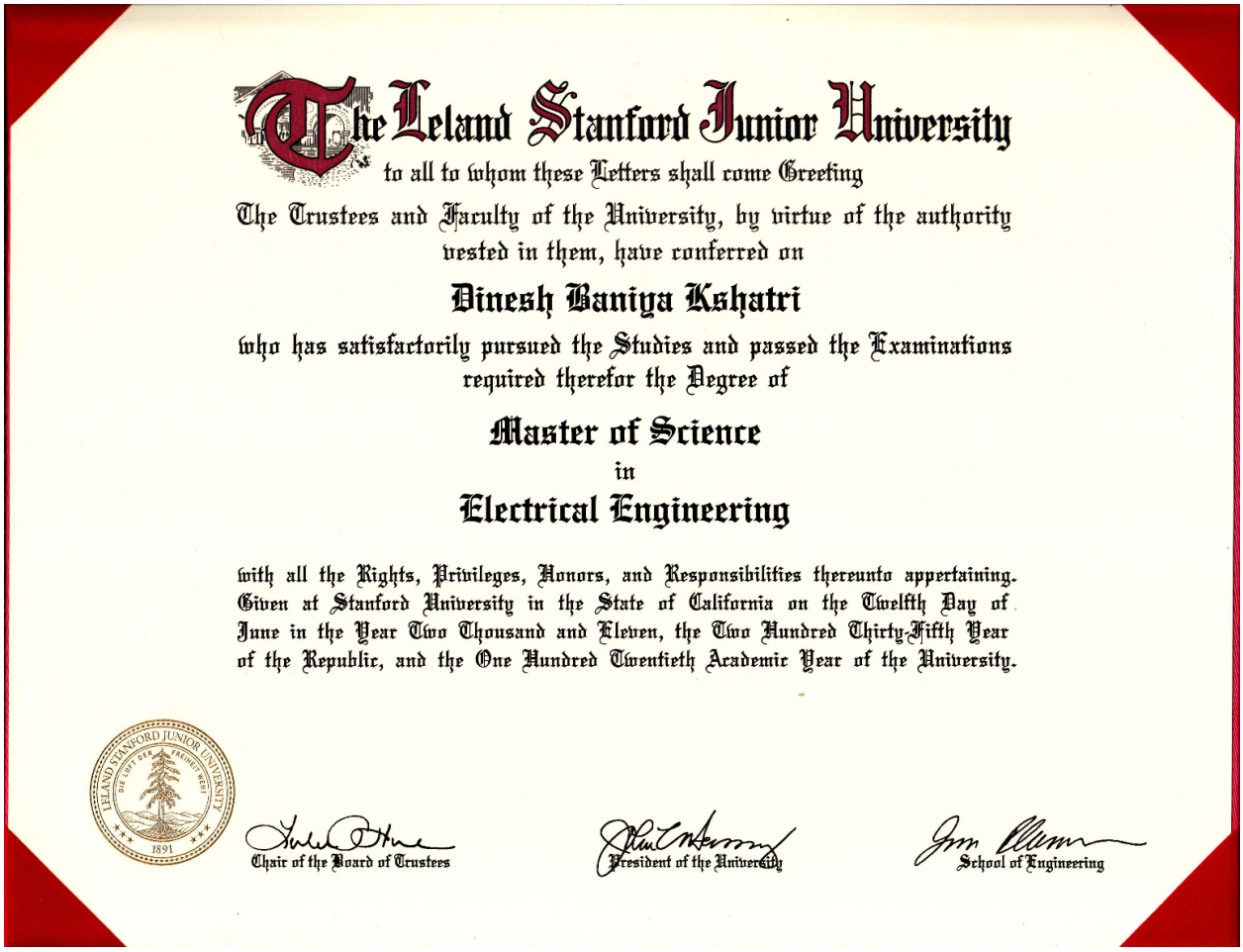
Mr. Abhinav Chalise, Mr. Nimesh Gopal Pradhan, Mr. Nishan Khanal, and Mr. Prashant Raj Bista are affiliated as Undergraduate Students in the Department of Electronics and Computer Engineering.

We fully support Mr. Abhinav Chalise, Mr. Nimesh Gopal Pradhan, Mr. Nishan Khanal, and Mr. Prashant Raj Bista in their application for the present research grant call of the National College of Engineering, Tribhuvan University, Nepal.

Kind regards,

Asst. Prof. Dr. Khem Gyanwali
Campus Chief

Appendix-3: Copy of Certificate of Highest Education Degree of Principal Investigator.



The Leland Stanford Junior University

to all to whom these Letters shall come Greeting

The Trustees and Faculty of the University, by virtue of the authority vested in them, have conferred on

Dinesh Baniya Kshatri

who has satisfactorily pursued the Studies and passed the Examinations required therefor the Degree of

Master of Science

in

Electrical Engineering

with all the Rights, Privileges, Honors, and Responsibilities thereunto appertaining. Given at Stanford University in the State of California on the Twelfth Day of June in the Year Two Thousand and Eleven, the Two Hundred Thirty-Fifth Year of the Republic, and the One Hundred Twentieth Academic Year of the University.

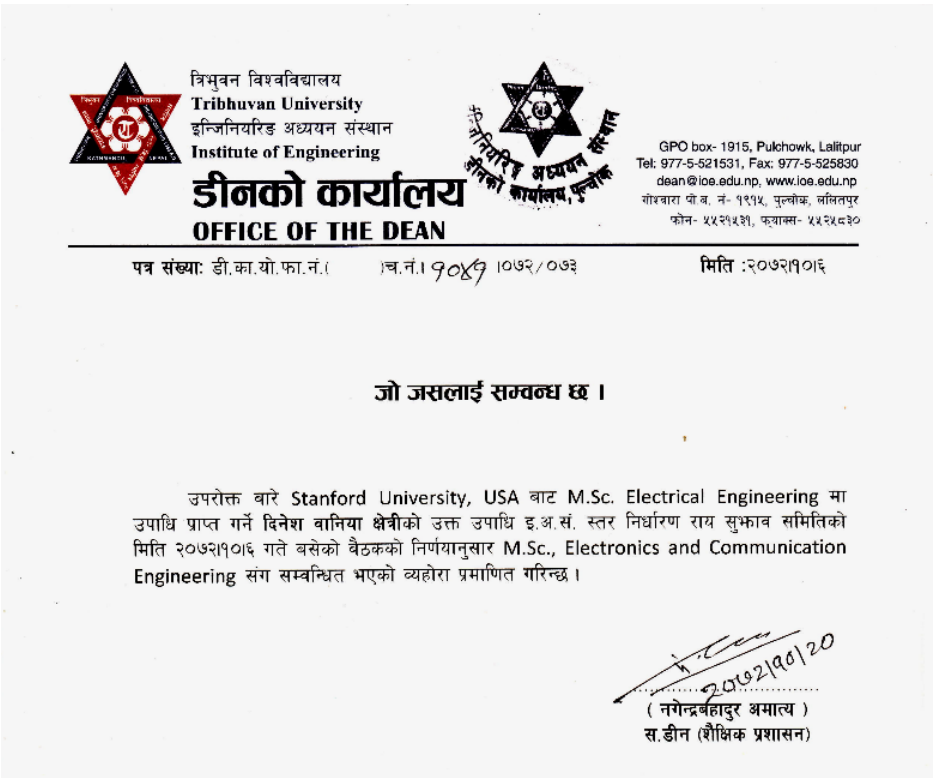
LELAND STANFORD JUNIOR UNIVERSITY
DIE LUFT DER FREIHEIT WEHT
1891

Chair of the Board of Trustees

President of the University

School of Engineering

Appendix-4: Equivalency of certificate belonging to principle investigator.



त्रिभुवन विश्वविद्यालय
Tribhuvan University
इन्जिनियरिङ अध्ययन संस्थान
Institute of Engineering

डीनको कार्यालय
OFFICE OF THE DEAN

GPO box- 1915, Pulchowk, Lalitpur
Tel: 977-5-521531, Fax: 977-5-525830
dean@ioe.edu.np, www.ioe.edu.np
गोश्वारा पो.ब. नं- १९१५, पुल्चोक, ललितपुर
फोन- ५५२१५३१, फ्याक्स- ५५२५८३०

पत्र संख्या: डी.का.यो.फा.नं.( )च.नं.। 9089 ।०७२/०७३ **मिति** :२०७२।१०।६

जो जसलाई सम्बन्ध छ ।

उपरोक्त बारे Stanford University, USA बाट M.Sc. Electrical Engineering मा उपाधि प्राप्त गर्ने **दिनेश वानिया क्षेत्री**को उक्त उपाधि इ.अ.सं. स्तर निर्धारण राय सुझाव समितिको मिति २०७२।१०।६ गते बसेको बैठकको निर्णयानुसार M.Sc., Electronics and Communication Engineering संग सम्बन्धित भएको व्यहोरा प्रमाणित गरिन्छ ।

2072/10/20

(नगेन्द्रबहादुर अमात्य)
स.डीन (शैक्षिक प्रशासन)

त्रिभुवन विश्वविद्यालय
Tribhuvan University
इन्जिनियरिङ अध्ययन संस्थान
Institute of Engineering

डीनको कार्यालय
OFFICE OF THE DEAN

GPO box- 1915, Pulchowk, Lalitpur
Tel: 977-5-521531, Fax: 977-5-525830
dean@ioe.edu.np, www.ioe.edu.np
गोश्वारा पो.ब. नं- १९१५, पुल्चोक, ललितपुर
फोन- ५५२१५३१, फ्याक्स- ५५२५८३०

पत्र संख्या: डी.का.यो.फा.नं.( )च.नं.। ८०४ ।०७३/०७४ **मिति** :२०७३।१०।१४

जो जसलाई सम्बन्ध छ ।

Stanford University, USA बाट श्री दिनेश वानिया क्षेत्रीले प्राप्त गर्नु भएको M.Sc. in Electrical Engineering उपाधि M.Sc. in Information and Communication Engineering संग सम्बन्धित भएको प्रमाणित गराई पाउँ भनी दिनु भएको निवेदन सम्बन्धमा इ.अ.सं. स्तर निर्धारण तथा राय सुझाव समितिको मिति २०७३।१०।१४ गते बसेको बैठकले निजको उक्त उपाधि Information and Communication Engineering संग सम्बन्धित भएको व्यहोरा प्रमाणित गरिन्छ ।

(नगेन्द्रबहादुर अमात्य)
स.डीन (शैक्षिक प्रशासन)

Appendix-5: Declaration from Principal Investigator.

त्रिभुवन विश्वविद्यालय
Tribhuvan University
इन्जिनियरिङ्ग अध्ययन संस्थान
Institute of Engineering
थापाथली क्याम्पस
Thapathali Campus
इलेक्ट्रोनिक्स तथा कम्प्युटर इन्जिनियरिङ्ग विभाग
Department of Electronics & Computer Engineering

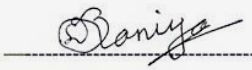
GPO Box- 280, Thapathali, Kathmandu
Tel: 4-246465, 4218300, Fax: 977-1-4247340
E-mail: doec@tcioe.edu.np
गोश्वारा पो. नं. २८०, थापाथली, काठमाडौं
फोन-४२४६४६५, ४२४६३०७
फ्याक्स: ५७७-१-४२४७३४०

Declaration

I hereby declare that all the statements and information given above are true.

If the Research Grant is awarded from the National College of Engineering (NCE), I will be in line with my duty as a Principal Investigator to conduct the proposed research project stated in this application and comply with Research grant rules and regulations.

The NCE has my permission to tangibly and electronically store the applicants' data which are required for the Research Grant application. If the Research Grant is awarded, the following information may be included in NCE's publications and website involved in science promotion and cooperation: surname, given name, academic degree, research field, place and name of my current employer, place, and name of my host institution, and the title, abstract, and keywords to define the research project.



Mr. Dinesh Baniya Kshatri
Assistant Professor
Electronics and Computer Engineering Department
Institute of Engineering, Thapathali Campus
Kathmandu, Nepal

November 26, 2023