# 

# 

# SCHOOL OF INFORMATION SCIENCES DEPARTMENT OF APPLIED INFORMATICS MSc IN ARTIFICIAL INTELLIGENCE AND DATA ANALYTICS

# Thessaloniki, 9/8/2025

# PROPOSAL FOR MSc THESIS

## 1. General

* **Student’s Full Name:** Vasileios-Efraim Tsavalias
* **Supervisor:** Prof. Apostolos Ampatzoglou
* **Research Field:** Applied Artificial Intelligence, Computer Vision, Digital Humanities, Natural Language Processing
* **Indicative Title:** Text-Guided Latent Diffusion for Semantic Inpainting of Ancient Greek Pottery
* **Keywords:**
  + **ACM Classification:**
  + Computing methodologies -> Artificial intelligence
  + Computing methodologies -> Computer vision
  + Applied computing -> Arts and humanities

**Free Keywords:**

* + - Diffusion Models
    - Image Inpainting
    - Cultural Heritage
    - Multimodal Learning
    - Ancient Greek Pottery

### 

### 2. Detailed Proposal Description

#### **A. Short description** (up to 200 words)

The digital restoration of damaged cultural heritage artifacts presents a persistent challenge for historical preservation. While established inpainting methods can fill missing regions based on visual cues like texture and color, they generally lack a semantic understanding of the subject. This limitation often produces reconstructions that are visually plausible but historically or iconographically inconsistent. This research addresses that gap by proposing a context-aware restoration framework. Our central hypothesis is that by conditioning a generative model on historical and mythological text descriptions, we can produce restorations that are not only visually coherent but also semantically and historically precise. The proposed methodology employs a Latent Diffusion Model, fine-tuned on a curated dataset of Greek pottery and guided by text embeddings from museum catalogs. The core contribution will be a quantitative and qualitative evaluation of the improvements offered by this text-guided approach over an unconditioned baseline.

#### **B. State of the Art** (up to 500 words)

Digital image restoration has traditionally relied on two main classes of inpainting algorithms: diffusion-based methods that propagate information from the boundary of a missing region inwards, and patch-based methods, that sample and stitch textures from known parts of the image. Although effective for repeating textures and minor gaps, these approaches tend to fail when a missing region requires a conceptual understanding of the image, as they are unable to generate novel, semantically coherent structures.

The emergence of deep generative models, particularly Generative Adversarial Networks (GANs), marked a notable advance in generating plausible image content. However, GANs are often difficult to train and can be prone to mode collapse. More recently, Denoising Diffusion Probabilistic Models (DDPMs), introduced by Ho et al. (2020), have become a leading approach for high-fidelity image synthesis. These models operate by progressively adding noise to an image and then learning the reverse denoising process, allowing for the generation of an image from a random noise vector.

The work of Rombach et al. (2022) on Latent Diffusion Models (LDMs) rendered this process more computationally efficient by applying the diffusion in a lower-dimensional latent space, which in turn enabled high-resolution image synthesis. This technology is the foundation for models such as Stable Diffusion. For the specific task of inpainting, Lugmayr et al. (2022) showed in “RePaint” how pre-trained DDPMs could be adapted to fill missing image regions without task-specific training. Building on this, the concept of guiding generative models with text, as explored by Bar-Tal et al. (2022) in “Text2LIVE,” has indicated the potential for fine-grained, semantic control over image editing. This project builds on these foundations, seeking to apply text-conditioned latent diffusion to the nuanced domain of cultural heritage, where semantic accuracy is of primary importance.

#### **C. Goals – Contribution** (up to 300 words)

The principal goal of this thesis is to develop and validate a system that uses textual information to improve the automated, semantic restoration of 2D images of ancient Greek pottery. The main scientific contribution will be a rigorous evaluation of how textual guidance affects the historical and iconographical accuracy of the final restoration when compared to purely visual methods.

The key research objectives are:

* To investigate whether a text-conditioned diffusion model can generate more historically and semantically accurate restorations of ancient pottery relative to an unconditioned baseline model.
* To establish a methodology for curating an effective multimodal (image-text) dataset for the specific domain of ancient Greek pottery, with a focus on addressing challenges in data quality and consistency.
* To quantify the impact of textual context on the inpainting process, using both standard image quality metrics (PSNR, SSIM) and qualitative analysis to assess the semantic integrity of the generated content.

#### **D. Methodology** (up to 1000 words)

The research methodology follows a structured, nine-stage computational pipeline designed for reproducibility and robustness. Each stage is an independent component that feeds into the next, ensuring a clear and traceable workflow from data acquisition to final model evaluation.

**Stage 1: Data Acquisition**

The foundational stage involves the programmatic collection of artifact images. A custom data acquisition component queries the Wikimedia Commons API, starting from a seed category (e.g., "Category:Ancient\_Greek\_pottery\_in\_the\_Louvre") and recursively traversing its subcategories to gather relevant image URLs. This process is subject to a configurable download limit to manage dataset size, resulting in a local collection of raw image files.

**Stage 2: Exploratory Data Analysis and Visualization (EDAV)**

Once the raw data is acquired, a comprehensive EDA is performed. This stage automatically analyzes the image collection to extract key metadata, such as image dimensions, file formats, and color modes. The output of this stage provides a clear statistical and visual overview of the dataset's characteristics (e.g., the generation of `summary\_statistics.txt` and plots like `file\_format\_distribution.png`), which is crucial for informing the subsequent processing steps.

**Stage 3: Data Processing**

To ensure consistency for the deep learning model, all raw images undergo a standardization process. Each image is resized to a uniform resolution of 512x512 pixels and converted to the PNG format. This step eliminates variability in input data and prevents potential errors during model training (e.g., processing `855 Jarra ática (51311273086).jpg` into a standardized PNG file).

**Stage 4: Data Splitting**

The curated and processed dataset is partitioned into three distinct subsets to ensure unbiased evaluation of the model. The data is divided into a training set (70%), a validation set (15%), and a final holdout test set (15%). This strict separation is critical for training the model, tuning its parameters, and assessing its generalization performance on unseen data.

**Stage 5: Feature Engineering (Masking)**

To simulate the artifact restoration task for the inpainting model, artificial damage is introduced into the images. For each image in the training, validation, and test sets, a corresponding mask is generated. This is achieved by creating random rectangular masks that obscure a portion of the image, thereby creating the ` (image, mask) ` pairs required for training the model to reconstruct the missing regions.

**Stage 6: Hyperparameter Tuning**

The selection of optimal model hyperparameters is a critical step for achieving peak performance. While this pipeline stage is designed to integrate with automated tuning libraries (e.g., Optuna, Ray Tune) to systematically search for the best combination of learning rate, batch size, and other parameters, the initial MVP run utilizes a set of predefined, validated hyperparameters to ensure a complete and rapid pipeline execution for baseline modeling.

**Stage 7: Model Training**

This stage involves fine-tuning a pre-trained diffusion model for the specific task of artifact inpainting. The core model, sourced from the Hugging Face model hub (e.g., `runwayml/stable-diffusion-inpainting`), is trained on the masked image dataset. The model learns to predict and reconstruct the missing pixels within the masked regions by leveraging the visual context of the surrounding, undamaged portions of the artifact.

**Stage 8: Model Evaluation**

The performance of the fine-tuned model is rigorously assessed using the unseen test set. The evaluation is both quantitative and qualitative. Quantitative analysis involves calculating objective metrics such as the Peak Signal-to-Noise Ratio (PSNR) and Structural Similarity Index (SSIM) to measure the pixel-level accuracy of the reconstructions. Qualitatively, the restored images are visually compared against their original, ground-truth counterparts to assess the perceptual quality and semantic correctness of the inpainting.

**Stage 9: Deployment Preparation**

The final stage of the pipeline involves packaging all necessary components for deployment or further use. This includes the fine-tuned model weights, the final hyperparameters file, and a README document. This creates a self-contained, portable package that allows the trained model to be easily shared, tested, or integrated into other applications.

### 3. Dissemination of Research Results

The findings from this research are intended for submission to reputable scientific conferences and journals. Potential venues include:

**Conferences:**

* [ECCV (VISART Workshop)](https://eccv.ecva.net/)
* [EuroVis](https://www.eurovis.org/)
* [CAA (Computer Applications and Quantitative Methods in Archaeology)](https://caa-international.org/conference/future-conferences/)

**Journals:**

* [International Journal of Computer Vision (IJCV)](https://www.springer.com/journal/11263)
* [ACM Journal on Computing and Cultural Heritage (JOCCH)](https://jocch.acm.org/)
* [Journal of Cultural Heritage](https://www.sciencedirect.com/journal/journal-of-cultural-heritage)
* [Digital Scholarship in the Humanities](https://academic.oup.com/dsh)
* [Computer Vision and Image Understanding](https://www.sciencedirect.com/journal/computer-vision-and-image-understanding)
* [Journal of Archaeological Science](https://www.sciencedirect.com/journal/journal-of-archaeological-science)
* [Heritage](https://www.mdpi.com/journal/heritage)

### 4. Bibliography / References

1. **Rombach, R., et al. (2022).** *High-Resolution Image Synthesis with Latent Diffusion Models*. <https://arxiv.org/abs/2112.10752>
2. **Ho, J., et al. (2020).** *Denoising Diffusion Probabilistic Models*. <https://arxiv.org/abs/2006.11239>
3. **Lugmayr, A., et al. (2022).** *RePaint: Inpainting using Denoising Diffusion Probabilistic Models*. <https://arxiv.org/abs/2201.09865>
4. **Bar-Tal, O., et al. (2022).** *Text2LIVE: Text-Driven Layered Image and Video Editing*. <https://arxiv.org/abs/2204.02491>
5. **Jaramillo, P., & Sipiran, I. (2024).** *Cultural Heritage 3D Reconstruction with Diffusion Networks*. [https://arxiv.org/abs/2410.10927](https://arxiv.org/abs/2410.10927%20)