# ATHENA: Archaeological Texture-aware Heritage Enhancement via Neural inpainting Architecture

## A Reproducible End-to-End Pipeline for Ancient Greek Pottery Restoration

Vasileios-Efraim Tsavalias[1], Christina Volioti[2], and Apostolos Ampatzoglou[1]

[1] Department of Applied Informatics, University of Macedonia, Thessaloniki, Greece
[2] Department of Education, School of Philosophy and Education, Aristotle University of Thessaloniki, Thessaloniki, Greece
aid25006@uom.edu.gr; chvolioti@gmail.com; a.ampatzoglou@uom.edu.gr

**Abstract.** Ancient Greek pottery is an important source of archaeological and historical evidence, as its painted scenes offer insight into artistic practice, mythology, ritual, warfare, and everyday life across the ancient Mediterranean. However, damage and missing surface areas often interrupt these scenes and their decorative patterns, making the original appearance of an object harder to study and interpret. The availability of open-access imagery makes Ancient Greek pottery a practical case study for controlled computational restoration research. Within this context, ATHENA was developed as a reproducible, end-to-end pipeline designed to generate restoration candidates and evaluate their visual fidelity, structural coherence, and usefulness for expert inspection. To build this pipeline, 4,110 source images yielded 5,553 object-centred crops. Eight inpainting methods were evaluated, alongside a structured study involving six specialists. Among the methods tested, fine-tuned Stable Diffusion with test-time augmentation (FT-SD+TTA) achieved the strongest quantitative performance, with the highest Peak Signal-to-Noise Ratio (17.37) and lowest Learned Perceptual Image Patch Similarity (0.392). By contrast, Large Mask Inpainting (LaMa) achieved the highest Structural Similarity Index Measure (0.468) and the largest expert preference share (36.7%). Overall, expert plausibility ratings averaged 3.53 out of 5. LaMa received 22 of 60 expert selections versus 20 for FT-SD+TTA; given the six-expert cohort and this narrow two-selection difference, the result points to only a small, descriptive preference for LaMa rather than a decisive one. Ultimately, ATHENA provides a proof-of-concept benchmark and a public implementation for comparing inpainting methods under a common data, masking, evaluation, and expert-review protocol.

**Keywords:** Image Inpainting; Cultural Heritage; Ancient Greek Pottery

## 1 Introduction

Painted Greek vessels document ancient artistic production, iconography, ritual, and social life [1]. Their shapes and decoration also support comparisons among painters,

regional styles, and changing visual themes [1]. However, abrasion, cracking, flaking, staining, deposits, and earlier interventions can obscure contours, gestures, ornaments, and surface relations on which interpretation depends [2]. Addressing such damage through physical conservation requires specialized expertise, time, and resources, so interventions must be accurately prioritized. Moreover, some treatments may be difficult to reverse without affecting the original materials [2]. By contrast, open-access collections allow damaged-vessel imagery to be studied at a scale that physical intervention alone cannot address. Digital inpainting can generate non-invasive visual hypotheses from the surviving context [3]-[5], but because the missing decoration cannot be known with certainty, each generated output should be treated as a candidate for inspection rather than as evidence of the original. Against this background, several recent heritage studies have applied generative methods [6]-[8] to produce coherent candidates. More specifically, Zhang [6] applied Stable Diffusion (SD) and LoRA to 47 Yangshao pottery images and evaluated image metrics and specialist criteria. Liao et al. [7] coupled adversarial prediction with reverse diffusion using eight ancient ceramic source images from plates augmented through rotations and synthetic damage, while Jiang et al. [8] trained prompt-guided residual diffusion on two mural datasets and ten degradation types. Although their results were informative within each setting, it should be emphasized that a visually convincing fill may alter an archaeologically important contour, ornament, or figure. Therefore, quantitative scores should be interpreted alongside specialist judgement [1], [2], [6]-[8].

Building on this, ATHENA evaluates eight methods spanning classical, deep-inpainting, and diffusion-based approaches on the same held-out crops and fixed masks, reducing variation caused by different datasets, test images, and damage definitions. A three-block expert study complements the quantitative evaluation and enables comparison between metric-based rankings and specialist preferences. Therefore, ATHENA contributes a controlled and reproducible comparison framework, rather than introducing another specialized restoration architecture. The proposed pipeline records data acquisition, filtering, source-grouped dataset splits, model configurations, generated outputs, statistics, and expert-study materials, while its public implementation supports reuse. These objectives lead to the following questions:

**RQ1.** Do classical, pre-trained deep, and diffusion-based inpainting methods preserve the visual and stylistic characteristics of Ancient Greek pottery under controlled synthetic damage?

*RQ1.1.* Which methods perform best across the selected quantitative metrics?

*RQ1.2.* How do domain experts assess the plausibility, perceived authenticity, and confidence of generated restoration candidates?

**RQ2.** What is the practical value of generated restoration candidates for expert-facing inspection?

*RQ2.1.* What does blinded pairwise evaluation indicate about their usefulness as inspection material or potential restoration guides?

*RQ2.2.* Which methods are preferred by experts in blinded model-comparison tasks?

*RQ2.3.* Do metric-based and expert-based rankings agree within the four-method expert-comparison subset?

## 2 Background Information

In this section the background information regarding inpainting research will be presented, including the following categories.

Classical propagation methods infer missing pixels from nearby visible information. More specifically, Bertalmio et al. [3] propagate image information from the mask boundary. In addition, the Navier-Stokes method [4] extends isophotes (lines of equal or similar intensity) across the gap through a fluid-flow analogy, whereas Telea [5] fills inward using nearby known pixels weighted by distance and boundary direction. These deterministic methods need no task-specific training, but local evidence can be insufficient for large or semantically complex missing regions [3]-[5].

Deep learning models use patterns acquired from training data. For example, Large Mask Inpainting (LaMa) [9] uses Fourier convolutions so distant regions and repeated patterns can guide the fill. Moreover, Mask-Aware Transformer (MAT) [10] retrieves relevant features while a dynamic mask excludes missing content. Co-Modulated Generative Adversarial Network (CoModGAN) [11] combines visible-image features with a random style representation and uses adversarial training to encourage realistic, varied completions. These capabilities may help with repeated ornament, but success on general images does not establish archaeological correctness [1].

Diffusion methods generate content through iterative denoising. One such method is latent diffusion [12] which works in a compressed representation and can use image or text conditions through cross-attention. Another method is RePaint [13] that repeatedly restores known pixels during reverse sampling.

## 3 Proposed Framework and Experimental Setup

The ATHENA pipeline contains 19 stages grouped into six modules. The 1st module establishes 1. reproducibility and 2. research design. The 2nd module includes 3. image acquisition, 4. unsuitable material filtering, and 5. pottery detections validation. The 3rd module generates 6. the classical baselines and conducts 7. exploratory analysis. The 4th module prepares 8. caption generation, 9. caption refinement, 10. processed images, 11. source-grouped splits, and 12. synthetic masks. Caption generation and refinement are used only for the SD prompt experiments, while the rest receive only the image crop and mask. The 5th module covers 13. tuning, 14. training, 15. baseline execution, and 16. evaluation, and finally, the 6th module prepares 17. model exports, 18. reports, and 19. expert-study materials. Figure 1 shows the image-level progression from a source image to a restoration candidate.

(a) Raw source (b) Object-centred crop (c) Synthetically damaged input (d) Restoration candidate

**Fig. 1.** Image-level progression through ATHENA from raw source to restoration candidate.

### 3.1 Dataset Construction

The corpus was assembled from openly licensed pottery images retrieved through Wikimedia Commons [17] and Europeana [18]. Acquisition collected 8,393 images, including 8,361 from Wikimedia Commons and 32 from Europeana. Accordingly, the corpus reflects material available online rather than a statistical census of surviving pottery. After color and integrity checks, 8,387 images entered filtering. Acceptance required visible pottery, adequate resolution and framing, sufficient object coverage, and acceptable sharpness. You Only Look Once version 8x (YOLOv8x) [19] supplied one bounding box for each accepted detection, and the pipeline padded and extracted each box separately. An image with five accepted vessels could therefore yield five object-centered crops. The Tenengrad focus measure [20] estimated sharpness from image gradients. In total, 4,110 retained source images produced 5,553 crops at 512 × 512 pixels (see Table 1).

**Table 1.** Dataset and evaluation construction summary.

| Component | Value |
|---|---|
| Raw acquisition | 8,393 images: 8,361 Wikimedia Commons; 32 Europeana |
| Filtering input | 8,387 images after color and integrity checks |
| Filtering rejections | 4,277 images: 2,587 no detection; 1,685 low confidence; 5 blur |
| Accepted source images | 4,110 filtered open-access pottery images |
| Accepted object crops | 5,553 object-centered 512 × 512 crops; one per accepted detection |
| Source-grouped split | 4,472 train / 523 validation / 558 test crops |
| Evaluation cases | 1,674 crop-mask pairs from 558 test crops × 3 masks |
| Mask families | Edge, irregular, rectangular |
| Mean whole-image mask coverage | 18.3% |

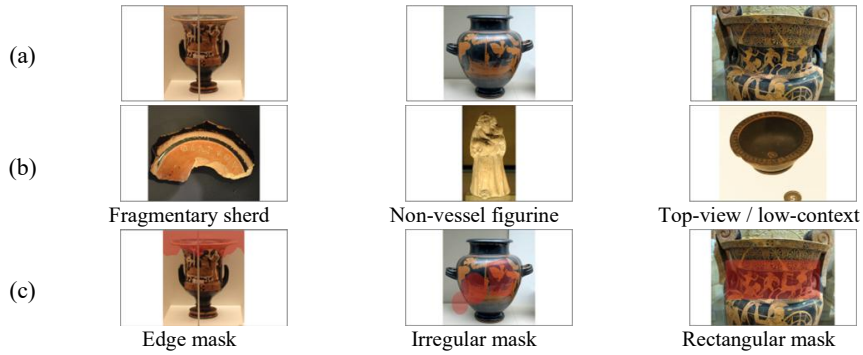


**Fig. 2.** (a) Accepted images, (b) rejected images and (c) synthetic damage with the three mask families.

To prevent leakage between partitions, crops were split by acquired-image identifier so that all crops from one image remained together. A post-split check found no overlap among training, validation, and test source IDs. The resulting split contained 4,472 training, 523 validation, and 558 test crops.

Because the unmasked crop remained available as a reference, synthetic damage enabled paired evaluation. Each test crop received one irregular, one rectangular, and one edge mask. Irregular masks combined curved strokes and blobs, rectangular masks used softened boxes, and edge masks entered from a crop boundary. Masks were restricted to an estimated foreground and regenerated when needed. The 558 test crops therefore yielded 1,674 crop-mask cases with mean whole-image coverage of 18.3%. Fig. 2 presents the accepted, rejected, and masked examples.

### 3.2 Restoration Methods

The benchmark compares eight methods (see Table 2) on identical crops and masks, grouped into three families: two classical methods, three deep-inpainting methods, and three SD configurations. The classical methods are Telea [5] and Navier–Stokes inpainting [4], whereas LaMa [9], MAT [10], and CoModGAN [11] provide the pretrained deep baselines. These five methods receive only the damaged crop and its binary mask and do not use text prompts or pottery-specific retraining. The remaining three methods are based on SD. Vanilla SD uses a publicly released inpainting model without pottery-specific retraining [12], [21]. Fine-tuned Stable Diffusion (FT-SD) adapts the U-Net to pottery while keeping the Variational Autoencoder (VAE) and text encoder fixed. FT-SD with test-time augmentation (FT-SD+TTA) processes a horizontally flipped copy, flips the output back, and averages the predictions.

**Table 2.** Compared inpainting methods and their role in the benchmark.

| Family | Method | Role |
|---|---|---|
| Classical | Telea | Fast-marching local baseline |
| Classical | Navier–Stokes | PDE-based local baseline |
| Deep zero-shot | LaMa | Image-wide Fourier-convolution model |
| Deep zero-shot | MAT | Dynamic-mask attention model |
| Deep zero-shot | CoModGAN | Co-modulated adversarial model |
| Diffusion | Vanilla SD | Public pretrained inpainting model |
| Diffusion | FT-SD (fine-tuned Stable Diffusion) | Pottery-adapted U-Net |
| Diffusion | FT-SD+TTA (test-time augmentation) | FT-SD with two-pass averaging |

SD first encodes the damaged image into a compressed latent representation. The U-Net then iteratively denoises this representation using the visible image, mask, and, where applicable, a text prompt. Finally, the VAE decodes the completed latent representation into an image [12]. Prompt ablation was conducted for Vanilla SD, FT-SD and FT-SD+TTA. Four prompt conditions were compared: an empty prompt, raw metadata, a natural-language caption generated by BLIP-2 [14], and a version of that caption shortened and reorganized by Qwen2.5 to fit the text encoder's token limit [15]. Telea, Navier–Stokes, LaMa, MAT, and CoModGAN were not included in the prompt ablation because they do not accept textual input; these methods received only the damaged crop and its binary mask.

All diffusion runs used 50 denoising steps and a guidance scale of 7.5, the documented defaults of the selected pipeline [21]. The step count controls iterative

refinement, while the guidance scale controls how strongly text steers the output [22], [21]. For reproducibility, seed 42 was reused across comparable runs to initialize the pseudo-random generation process. Reusing the same seed reduces one source of stochastic variation across comparable runs, although exact reproduction also depends on the documented software, library versions, and hardware environment [16].

### 3.3 Quantitative Evaluation

ATHENA interprets complementary metrics separately rather than combining them into one score. More specifically, these metrics are: (a) Peak Signal-to-Noise Ratio (PSNR) is derived from mean squared pixel error, where higher values indicate closer numerical agreement; (b) Structural Similarity Index Measure (SSIM) [23] compares local brightness, contrast, and spatial intensity patterns, so higher values indicate closer edges, contours, and texture organization; (c) Learned Perceptual Image Patch Similarity (LPIPS) [24] compares neural features; lower values indicate greater generic perceptual similarity, not archaeological correctness; (d) Fréchet Inception Distance (FID) [25] and (e) Kernel Inception Distance (KID) [26] compare feature distributions across restored and reference sets rather than individual motifs; (f) COLOR compares CIE-Lab lightness and color distributions [27] through weighted Wasserstein distances [28], where higher values indicate closer color distributions, not correct spatial placement; and finally, (g) PATTERN uses Local Binary Patterns [29] and histogram intersection to compare local micro-texture frequencies, not larger ornaments or contours. Primary PSNR, SSIM, LPIPS, COLOR, and PATTERN scores were computed over each mask's bounding box. Full-image versions were reported as supplementary measures. LPIPS used AlexNet [24], while full-image FID [25] and KID [26] were treated as secondary dataset-level diagnostics.

Because every method and prompt condition used the same crop-mask cases, matching sample identifiers produced paired observations. Pairwise differences were analyzed using paired t-tests, 95% confidence intervals, and paired Cohen's d following Lakens [30]. Wilcoxon signed-rank tests [31] provided a non-parametric check, while Bonferroni correction [32] controlled false-positive risk across multiple comparisons. Since three masks came from each crop, inference remains at crop-mask level.

### 3.4 Expert Evaluation

As already mentioned, in addition to the quantitative evaluation, an expert evaluation was also conducted, in which six specialists participated: two archaeologists, one museologist, one philologist trained in history and archaeology, one philologist, and one philologist-educator. The expert study was analyzed separately from the metric benchmark and included three blocks (see Fig. 3). Block A collected five-point ratings for perceived authenticity, archaeological plausibility, and confidence, together with a mandatory comment. Block B presented an undamaged reference and a generated candidate in blinded A/B order. Block C presented four methods with randomized, concealed A–D labels. These four methods were the strongest diffusion configuration, FT-SD+TTA, and three deep-inpainting baselines, LaMa, MAT, and CoModGAN.

Participants completed on-screen consent, exports used anonymized identifiers, and the platform recorded 210 responses with mandatory comments. Comments were coded

and grouped into recurring themes, with illustrative quotations selected following qualitative content-analysis principles [33].

The public ATHENA interface administered the expert evaluation, while a protected administration dashboard was used to configure the campaign, monitor completion and quality flags, and export anonymized CSV and JSON responses[1].

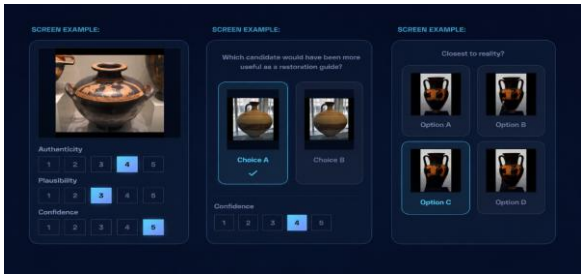


**Fig. 3.** Interfaces for Blocks A, B, and C of the expert study.

# 4 Results

## 4.1 RQ1. Preservation of Visual and Stylistic Characteristics

RQ1 is evaluated through the quantitative benchmark and Block A. Specifically, RQ1.1 uses the aggregate quantitative benchmark, whereas RQ1.2 uses plausibility, perceived authenticity, confidence, and comment data (Block A).

Table 3 presents the common no-text benchmark across all eight methods. Telea, Navier–Stokes, LaMa, MAT, and CoModGAN received only the damaged crop and mask, as these methods do not accept textual input. Vanilla SD, FT-SD, and FT-SD+TTA were evaluated using an empty prompt in this common benchmark. Additional prompt-conditioned experiments using raw metadata, BLIP-2-generated captions, and Qwen2.5-refined captions were conducted separately for all three SD configurations. Because these prompt conditions did not improve PSNR relative to the corresponding empty-prompt runs, their results are reported separately rather than included in Table 3.

Across the common benchmark, FT-SD+TTA showed the strongest overall performance resulting in the highest PSNR, lowest LPIPS, lowest FID, and highest PATTERN score. It also achieved the second-highest SSIM score, behind LaMa, which obtained the highest SSIM score. Vanilla SD and FT-SD produced almost identical results, achieving the lowest KID score and the highest COLOR score. MAT achieved the second-highest COLOR score, but ranked below LaMa and the SD variants on most other measures. CoModGAN generally performed below the other learned methods. Finally, Telea and Navier–Stokes produced the lowest PSNR values, approximately four to five PSNR points below the strongest learned methods.

Within the diffusion family, FT-SD differed from Vanilla SD by -0.0069 in PSNR and -0.0004 in SSIM. Although the corresponding paired Cohen's d values were -0.46

[1] ATHENA interfaces for participant: https://athena-evaluation.vercel.app/; for administration: https://athena-evaluation.vercel.app/admin.

and -0.54, indicating moderate standardized differences, the changes in the original metric units were very small. The LPIPS difference was not significant after Bonferroni correction. By contrast, FT-SD+TTA increased PSNR by 0.89 and SSIM by 0.028 and reduced LPIPS by 0.009 relative to FT-SD; all three changes were favorable because lower LPIPS is better. The corresponding Cohen's d values exceeded 1.0, indicating large paired standardized differences, although these results should not be interpreted as evidence of archaeological superiority.

Prompt conditioning did not improve PSNR for the tested SD variants. Relative to the corresponding empty-prompt runs, raw metadata, BLIP-2-generated captions [14], and Qwen2.5-refined captions [15] reduced PSNR by 0.53–1.03 under seed 42. These findings are therefore specific to the tested prompts, models, and fixed random initialization.

**Table 3.** Aggregate benchmark on 1,674 crop-mask pairs. Higher is better for PSNR, SSIM, COLOR, and PATTERN; lower is better for LPIPS, FID, and KID.

| Method | PSNR | SSIM | LPIPS | FID | KID | COL. | PAT. |
|---|---|---|---|---|---|---|---|
| Telea | 12.46 | 0.391 | 0.663 | 55.61 | 0.0925 | 0.087 | 0.820 |
| Nav.–Stokes | 12.92 | 0.405 | 0.665 | 53.54 | 0.0920 | 0.094 | 0.486 |
| LaMa | 16.60 | **0.468** | 0.466 | 22.69 | 0.0662 | 0.123 | 0.937 |
| MAT | 15.52 | 0.396 | 0.465 | 28.36 | 0.0706 | 0.148 | 0.914 |
| CoModGAN | 14.21 | 0.366 | 0.506 | 34.61 | 0.0740 | 0.114 | 0.897 |
| Vanilla SD | 16.48 | 0.399 | 0.401 | 16.83 | **0.0626** | **0.161** | 0.936 |
| FT-SD | 16.47 | 0.399 | 0.402 | 16.84 | **0.0626** | **0.161** | 0.937 |
| FT-SD+TTA | **17.37** | 0.427 | **0.392** | **16.56** | 0.0627 | 0.142 | **0.950** |

In addition, Block A addresses RQ1.2. Across 60 trials, mean plausibility was 3.53/5, mean perceived authenticity 3.15/5, and mean confidence 3.50/5. The mean ratings indicated near-midpoint perceived authenticity and moderately positive plausibility and confidence. However, because the evaluations were provided by only six specialists, these findings should be interpreted as preliminary evidence from the current cohort rather than as representative estimates of wider expert opinion.

### 4.2 RQ2. Practical Value for Expert Inspection

RQ2 is evaluated through the Block B pairwise task, the Block C four-method choice, and the comparison between the Block C expert-preference and metric-based rankings. These analyses address RQ2.1, RQ2.2, and RQ2.3, respectively.

Block B was evaluated using a predefined Tier-1 criterion. Practice trials were used for familiarization, while anchor trials supported calibration or attention checking. Practice trials, anchor trials, and Tie/Unsure responses were excluded from the denominator used to calculate the generated-candidate preference rate. Tier 1 required generated-candidate preference in at least 40% of the eligible trials. The aggregate rate was 0.234, and no participant met Tier 1. Accordingly, the evidence favors inspection and comparison over direct restoration-guide use.

Block C measured preference within the four-method subset across 60 trials. Figure 4 presents one representative held-out case and the four corresponding restoration candidates. LaMa received 22 of 60 selections (36.7%), FT-SD+TTA 20 (33.3%), MAT 12 (20.0%), and CoModGAN 6 (10.0%), as shown in Fig. 5. The expert-preference ranking was therefore LaMa, FT-SD+TTA, MAT, and CoModGAN, whereas PSNR and LPIPS placed FT-SD+TTA first. However, the 60 selections were repeated responses from six specialists rather than 60 independent reviewers, and the leading margin was only two selections. The preference for LaMa is therefore descriptive rather than conclusive.

(a) Damaged input (b) LaMa (c) FT-SD+TTA (d) CoModGAN (e) MAT

**Fig. 4.** One held-out damaged crop and the four Block C restoration candidates.

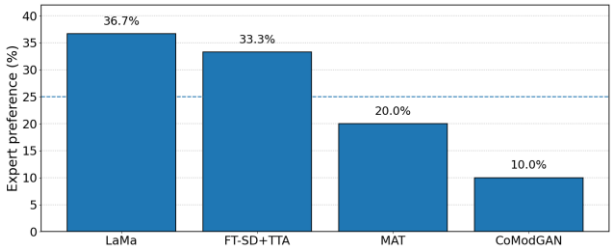


**Fig. 5.** Expert preference shares in Block C. The dashed line marks the uniform-choice baseline of 25%, expected if the four methods were selected equally often.

All 210 item-level responses included a comment. Recurring themes concerned practical usefulness, iconographic preservation, geometry, color, and the natural appearance of wear. Illustrative anonymized comments included: "*Most restoration candidates were suitable for teaching, with only minor losses of realism in surface wear and color*" and "*The iconographic composition was generally rendered satisfactorily, with minor losses in color intensity and the natural appearance of wear*".

## 5 Discussion

### 5.1 RQ1. Preservation of Visual and Stylistic Characteristics

**RQ1.** The evaluated methods preserved visual and stylistic information, but the result was not uniform across methods or criteria. FT-SD+TTA provided the strongest pixel-level and perceptual fidelity, while LaMa provided the strongest structural similarity. In addition, expert plausibility and perceived authenticity indicate moderate acceptance of the candidates for inspection rather than consistently convincing reconstruction.

**RQ1.1 quantitative preservation**. No method performed best across all quantitative measures, indicating that the metrics capture different aspects of

restoration quality. FT-SD+TTA achieved the strongest overall quantitative performance, leading PSNR, LPIPS, FID, and PATTERN, whereas LaMa achieved the highest SSIM. This suggests that FT-SD+TTA was more successful in reproducing pixel-level, perceptual, distributional, and local-texture characteristics, while LaMa better preserved local structural organization. Telea and Navier–Stokes obtained lower PSNR values, indicating that methods relying primarily on neighboring visible pixels may be less effective when damage interrupts a larger or semantically complex motif [3]-[5]. This is consistent with Liao et al. [7], who also reported lower reconstruction errors for a learned ceramic-restoration model than for classical baselines. However, the present findings are limited to the tested synthetic masks and do not imply that classical methods are unsuitable for all forms of damage.

Within the diffusion family, pottery-specific fine-tuning alone offered little improvement over Vanilla SD, while test-time augmentation produced clearer gains. Zhang [6] reported improvements from LoRA adaptation on a focused Yangshao pottery corpus, but differences in adaptation strategy, datasets, masks, and evaluation protocols prevent direct comparison. ATHENA's limited fine-tuning gains may indicate either that the pretrained model already captured sufficient local texture information or that the corpus was too heterogeneous for consistent adaptation. Both explanations remain tentative.

Prompt-conditioning results were also configuration-specific. Unlike Jiang et al. [8], who used domain-specific prompts and a dedicated fusion architecture, ATHENA supplied metadata or automatically generated descriptions to otherwise unchanged SD pipelines. The lower PSNR values therefore show only that the tested metadata, BLIP-2 captions, and Qwen2.5-refined captions did not improve Vanilla SD, FT-SD, or FT-SD+TTA under the selected settings and fixed seed.

**RQ1.2 expert evaluation**. Block A indicated moderate expert acceptance. Specialists identified potential value for teaching, museum display, and comparative inspection, while noting limitations in color, wear, and surface detail. Given the six-specialist cohort, these findings remain preliminary and specific to the evaluated cases.

### 5.2 RQ2. Practical Value and Metric–Expert Agreement

**RQ2.** Overall, the findings provide stronger support for using the generated candidates as material for expert comparison and inspection than as reliable restoration guides.

**RQ2.1 inspection and guide use**. In Block B, the generated-candidate preference rate remained below the predefined Tier-1 threshold, and no participant met the criterion. This may partly reflect the difficulty of assessing a completion when the missing motif cannot be verified, as well as the small expert cohort. The evidence therefore supports the use of generated candidates as visual hypotheses for comparative inspection rather than as direct guides for physical restoration.

**RQ2.2 preferred methods**. In the four-method comparison, LaMa received the largest preference share, with FT-SD+TTA only two selections behind. However, the 60 selections represented repeated responses from six specialists rather than judgements from 60 independent experts. The result should therefore be interpreted as a descriptive preference within this cohort and task, not as evidence that LaMa is generally superior.

**RQ2.3 ranking agreement**. Agreement between metric-based and expert-based rankings was partial. LaMa ranked first in expert preference and SSIM, whereas FT-SD+TTA ranked first in PSNR and LPIPS. This divergence is plausible because SSIM emphasizes local structural similarity and LPIPS generic feature-space similarity, while experts also considered iconographic motifs, surface wear, stylistic coherence, and the natural appearance of the completion. LaMa's broad-context processing [9], its leading SSIM score, and the particular cases included in Block C may have contributed to its preference advantage; however, the present data do not allow these influences to be isolated.

## 6 Conclusions

ATHENA provides a controlled benchmark of eight inpainting methods evaluated on identical crop-mask cases and complemented by expert evaluation. No method performed best across all criteria: FT-SD+TTA led PSNR, LPIPS, FID, and PATTERN, whereas LaMa achieved the highest SSIM and expert preference. Fine-tuning alone offered little improvement over Vanilla SD, while test-time augmentation produced clearer gains. Prompt conditioning did not improve PSNR across the three SD configurations under the tested conditions. Expert findings support using restoration candidates as visual hypotheses for inspection rather than reliable restoration guides. Despite limitations, ATHENA offers a reproducible framework for comparing restoration candidates under expert review and reuse.

**Acknowledgments**. This work was conducted within the MSc in Artificial Intelligence and Data Analytics at the University of Macedonia. Presentation funding was provided by the University of Macedonia Research Committee.

**Disclosure of Interests.** The authors have no competing interests to declare that are relevant to the content of this article.

## References

1. Oakley, J.H.: Greek Vase Painting. American Journal of Archaeology 113(4), 599–627 (2009).
2. El Ouahabi, M., Cools, C., Rousseau, V., Gautier, J.: Different cleaning techniques for archeological ceramics: A review. Heritage 8(10), 434 (2025).
3. Bertalmio, M., Sapiro, G., Caselles, V., Ballester, C.: Image inpainting. In: Proceedings of SIGGRAPH, pp. 417–424 (2000).
4. Bertalmío, M., Bertozzi, A.L., Sapiro, G.: Navier–Stokes, fluid dynamics, and image and video inpainting. In: CVPR, vol. 1, pp. I355–I362 (2001).
5. Telea, A.: An image inpainting technique based on the fast marching method. Journal of Graphics Tools 9(1), 23–34 (2004).
6. Zhang, X.: AI-assisted restoration of Yangshao painted pottery using LoRA and Stable Diffusion. Heritage 7(11), 6282–6309 (2024).
7. Liao, D., Zeng, T., Yang, J., et al.: Digital prediction of ancient ceramic images missing areas based on deep adversarial and reverse diffusion. npj Heritage Science 13, 242 (2025).

8. Jiang, C., Ren, T., Cheng, Z.: All-in-one mural restoration with prompt-guided residual diffusion. npj Heritage Science 13, 667 (2025).
9. Suvorov, R., Logacheva, E., Mashikhin, A., et al.: Resolution-robust large mask inpainting with Fourier convolutions. In: WACV, pp. 2149–2159 (2022).
10. Li, W., Lin, Z., Zhou, K., et al.: MAT: Mask-aware transformer for large hole image inpainting. In: CVPR, pp. 10758–10768 (2022).
11. Zhao, S., Cui, J., Sheng, Y., et al.: Large scale image completion via co-modulated generative adversarial networks. In: ICLR (2021).
12. Rombach, R., Blattmann, A., Lorenz, D., et al.: High-resolution image synthesis with latent diffusion models. In: CVPR, pp. 10684–10695 (2022).
13. Lugmayr, A., Danelljan, M., Romero, A., et al.: RePaint: Inpainting using denoising diffusion probabilistic models. In: CVPR, pp. 11461–11471 (2022).
14. Li, J., Li, D., Savarese, S., et al.: BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In: ICML, PMLR 202, 19730–19742, (2023).
15. Qwen Team: Qwen2.5 Technical Report. arXiv:2412.15115 (2024).
16. Pineau, J., Vincent-Lamarre, P., Sinha, K., et al.: Improving reproducibility in machine learning research: A report from the NeurIPS 2019 reproducibility program. Journal of Machine Learning Research 22(164), 1–20 (2021).
17. Wikimedia Commons. https://commons.wikimedia.org/.
18. Europeana. https://www.europeana.eu/.
19. Ultralytics YOLOv8. https://github.com/ultralytics/ultralytics.
20. Pech-Pacheco, J.L., Cristobal, G., Chamorro-Martinez, J., et al.: Diatom autofocusing in brightfield microscopy: A comparative study. In: ICPR, pp. 314–317 (2000).
21. Hugging Face Diffusers: Stable Diffusion inpainting. https://huggingface.co/docs/diffusers (accessed 2026/07/24).
22. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv:2207.12598 (2022).
23. Wang, Z., Bovik, A.C., Sheikh, H.R., et al.: Image quality assessment: From error visibility to structural similarity. IEEE Transactions on Image Processing 13(4), 600–612 (2004).
24. Zhang, R., Isola, P., Efros, A.A., et al.: The unreasonable effectiveness of deep features as a perceptual metric. In: CVPR, pp. 586–595 (2018).
25. Heusel, M., Ramsauer, H., Unterthiner, T., et al.: GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In: NeurIPS, vol. 30 (2017).
26. Binkowski, M., Sutherland, D.J., Arbel, M., Gretton, A.: Demystifying MMD GANs. In: ICLR (2018).
27. CIE: CIELAB colour space. https://cie.co.at/.
28. Rubner, Y., Tomasi, C., Guibas, L.J.: The Earth Mover's Distance as a metric for image retrieval. International Journal of Computer Vision 40, 99–121 (2000).
29. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(7), 971–987 (2002).
30. Lakens, D.: Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for t-tests and ANOVAs. Frontiers in Psychology 4, 863 (2013).
31. Wilcoxon, F.: Individual comparisons by ranking methods. Biomet. Bull. 1(6), 80–83 (1945).
32. Dunn, O.J.: Multiple comparisons among means. Journal of the American Statistical Association 56(293), 52–64 (1961).
33. Elo, S., Kääriäinen, M., Kanste, O., et al.: Qualitative content analysis: A focus on trustworthiness. SAGE Open 4(1), 1–10 (2014).