Assessing the Viability of Image-to-Image Models for Restoring Historical Photographs of News Articles

Lucien Heitz[1,*], Bruno N. Sotic[2], Qi Bi[3] and Jaap Kamps[2]

[1]University of Zurich, Zurich, Switzerland
[2]University of Amsterdam, Amsterdam, the Netherlands
[3]Utrecht University, Utrecht, the Netherlands

Abstract

This paper focuses on photorealistic restoration of historical newspaper images in the context of the MediaEval 2026 NewsImages challenge. We evaluate three open-weight diffusion-based image-editing frameworks: Flux.2 Klein, Qwen-Image-Edit-2509, and FireRed-Image-Edit-1.1. We assess model performance through complementary no-reference perceptual quality metrics, prompt-image alignment measures, and online user studies with both experts and non-experts. Results reveal a clear trade-off across evaluation dimensions: FireRed achieves the strongest per-image perceptual quality, and Flux shows the strongest prompt-level alignment in the offline assessment. In the online studies, Qwen is perceived as providing the highest title-image fit by both experts and non-experts, even outperforming originals. These findings highlight the potential of image restoration for providing news visuals. But they also reveal a mismatch between perceptual offline metrics and human feedback.

1. Introduction and Background

This paper contributes to the NewsImages 2026 Challenge [1] at MediaEval 2026,[1] which continues the multi-year benchmarking effort [2] for automated matching of articles with appropriate visuals. This year's iteration extends the previous comparison of retrieval- and generation-based pipelines by including a mix of both contemporary articles and scans of 19th- and 20th-century newspapers. As a *Quest for Insight* paper written by NewsImages organizers, this work focuses on the historical articles of the challenge and how their restoration affects perceived image fit. The images in the historical portion of the test set consist of scans of prints, where originals are typically low-resolution, monochromatic, and physically degraded.

Two findings from prior iterations of the benchmark motivate our work. First, a previous user study showed that AI-generated images were frequently perceived as a more fitting accompaniment to a news article than the editorially-assigned original [3]. Second, end-to-end generative pipelines were reported to outperform retrieval over the *Yahoo-Flickr Creative Commons 100 Million* dataset (YFCC100M),[2] even when retrieval was augmented with engineered prompts and modern vision-language encoders [4]. Together, these results suggest that automated visual augmentation of news content is both technically feasible and, at times, preferred by readers. We want to empirically verify if this also holds true for restored images.

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online
*Corresponding author.
heitz@ifi.uzh.ch (L. Heitz); nadalic.sotic@uva.nl (B. N. Sotic); q.bi@uu.nl (Q. Bi); kamps@uva.nl (J. Kamps)
0000-0001-7987-8446 (L. Heitz); 0009-0003-2122-2235 (B. N. Sotic); 0000-0002-1047-4790 (Q. Bi); 0000-0002-6614-0087 (J. Kamps)

[1]https://multimediaeval.github.io/editions/2026
[2]https://multimediacommons.org

2. Method

We compare three open-weight diffusion-based vision language models (VLMs) —Flux.2 Klein [5], Qwen-Image-Edit-2509 [6], and FireRed-Image-Edit-1.1 [7]—and compare the generated image with the editorially-selected originals.[3] We share all originals and restored images in our GitHub repository.[4] The PNG files from the models include the workflow used to generate them; they can be dragged and dropped into Comfy UI[5] to reproduce our results. The input was the original image with a model-specific restoration prompt.

Flux.2 Klein For the Flux.2 Klein [5],[6] we use the 4B base image-editing configuration. The text encoder is `qwen_3_4b.safetensors`. The VAE is `flux2-vae.safetensors`. The diffusion model is `flux-2-klein-base-4b-fp8.safetensors`, corresponding to the undistilled 4B base variant of Flux.2 Klein [5]. We use the original scan with the text prompt "*restore and recolor this old photograph, add camera imperfections and washed out colors, dull skin and not glossy.*" The goal was to preserve the visuals' historic character and implicitly convey to the reader that this is not about a recent event.

Qwen-Image-Edit-2509 For Qwen-Image-Edit-2509 [6][7], we used the diffusion model `qwen_image_edit_2509_fp8_e4m3fn.safetensors` in combination with the text encoder `qwen_2.5_vl_7b_fp8_scaled.safetensors` and the VAE `qwen_image_vae.safetensors`. We use the original scan with the text prompt again, "*restore and recolor this old photograph.*" Here, our goal was to restore the original scan so it looks like a photograph that was taken recently.

FireRed-Image-Edit-1.1 For FireRed-Image-Edit-1.1 [7],[8] we made use of the diffusion model `FireRed-Image-Edit-1.1-transformer.safetensors`, the main FireRed-Image-Edit-1.1 [7] transformer checkpoint for instruction-based image editing. The text encoder is `qwen2.5vl-7b-bf16.safetensors`, and the VAE is `qwen_image_vae.safetensors`. In this approach, we combined this with a dedicated `FireRed-Image-Edit-1.1-Lightning-8steps-v1.2.safetensors` LoRA. The image and text inputs are the same as with the previous model. Our goal was again to create a photograph that looks modern. Given that FireRed is half a year newer than Qwen, we can track the rapid progress of VLMs in image restoration capabilities.

3. Results and Analysis

We evaluate the restored images in both offline and online settings. In the offline setting, we conduct a quantitative comparison with reference benchmark metrics. The online setting consists of a user study with experts (participants of the NewsImages 2026 challenge, $N = 18$) and non-experts recruited on Prolific ($N = 30$). Both settings complement each other and allow us to examine how theoretical gains (assessed via offline metrics) translate into and affect perceived article-image fit (as reflected in user survey feedback).[9]

[3] This baseline consists of scans of monochromatic photographs from newspapers.
[4] https://github.com/Informfully/Challenges/tree/main/newsimages26/workflows/NewsImages
[5] https://comfy.org
[6] https://docs.comfy.org/tutorials/flux/flux-2-klein
[7] https://huggingface.co/Comfy-Org/Qwen-Image-Edit_ComfyUI/tree/main
[8] https://huggingface.co/FireRedTeam/FireRed-Image-Edit-1.1-ComfyUI
[9] The offline evaluation is performed on a corpus of 50 articles. The online evaluation, in contrast, uses only a subset.

Figure 1: Qualitative comparison between Flux.2 Klein [5], Qwen-Image-Edit-2509 [6], and FireRed-Image-Edit-1.1 [7].

3.1. Offline Evaluation

Since paired ground-truth restored images are unavailable, no-reference image quality assessment is conducted to measure perceptual naturalness and restoration quality. We compute the no-reference perceptual quality metrics [8], FID [9] and CLIP scores [10]. [10]

Table 1 and Table 2 summarizes the main quantitative comparison among Qwen-Image-Edit-2509, FireRed-Image-Edit-1.1, and Flux.2 Klein on the restoration task. FireRed-Image-Edit-1.1 achieves the best score on most no-reference perceptual quality metrics. It obtains the lowest NIQE, BRISQUE, and PI scores, indicating stronger natural-image statistics and lower perceptual distortion. It also achieves the highest NRQM, MUSIQ, MANIQA, CLIP-IQA, CLIP-IQA+, DB-CNN, NIMA, HyperIQA, PaQ-2-PiQ, LIQE, CNNIQA, and TReS scores, suggesting that its restored images are generally preferred by learned NR-IQA models.

These results indicate that FireRed-Image-Edit-1.1 produces visually natural and perceptually stable restorations under the no-reference evaluation setting. Flux.2 Klein achieves the highest CLIPScore and PickScore, suggesting stronger alignment with the editing prompt, but its no-reference perceptual quality scores are generally weaker than FireRed-Image-Edit-1.1. Overall, the results show a clear trade-off and model-specific strenghts. FireRed-Image-Edit-1.1 performs best in per-image perceptual quality, Qwen-Image-Edit-2509 best matches the real-image distribution, and Flux.2 Klein shows the strongest prompt-level alignment.

Overall, the restoration results suggest that no single framework dominates all evaluation dimensions. FireRed-Image-Edit-1.1 is the strongest model in per-image perceptual quality assessment, Qwen-Image-Edit-2509 is the strongest in distributional similarity to real old photographs, and Flux.2 Klein is the strongest model in prompt-level instruction alignment.

[10] Specifically, we report NIQE, BRISQUE, and PI as distortion-oriented metrics, where lower values indicate better perceptual quality. We further report learned NR-IQA and aesthetic metrics, including NRQM, MUSIQ, MANIQA, CLIP-IQA, CLIP-IQA+, DB-CNN, NIMA, HyperIQA, PaQ-2-PiQ, LIQE, CNNIQA, and TReS, where higher values generally indicate better perceptual quality. We also compute prompt-image alignment metrics, including CLIPScore and PickScore, where higher values indicate better alignment with the restoration instruction.

Table 1
Quantitative comparison on the restoration task. The results are averaged over 50 generated images. For NIQE, BRISQUE, PI, FID, and KID, lower is better. For the remaining metrics, higher is better. Best results are highlighted in bold.

Method	NIQE↓	BRISQUE↓	PI↓	NRQM↑	MUSIQ↑	MANIQA↑	CLIP-IQA↑	NIMA↑
Flux.2 Klein [5]	5.597	37.154	5.133	**5.372**	31.562	0.219	0.317	4.080
Qwen-Image-Edit-2509 [6]	7.834	53.943	7.089	3.741	35.228	0.236	0.384	4.023
FireRed-Image-Edit-1.1 [7]	**4.756**	**24.161**	**4.815**	5.206	**63.969**	**0.319**	**0.486**	**5.025**

Table 2
Additional no-reference perceptual and prompt-alignment metrics. Higher values indicate better performance. The results are averaged over 50 generated images. Best results are highlighted in bold.

Method	CLIP-IQA+↑	DB-CNN↑	HyperIQA↑	PaQ-2-PiQ↑	LIQE↑	CNNIQA↑	TReS↑	CLIPScore↑	PickScore↑
Flux.2 Klein [5]	0.319	0.368	0.308	61.508	1.133	0.492	33.334	**17.398**	**0.1863**
Qwen-Image-Edit-2509 [6]	0.414	0.358	0.280	59.719	1.316	0.336	31.710	12.146	0.1777
FireRed-Image-Edit-1.1 [7]	**0.557**	**0.477**	**0.495**	**71.296**	**3.323**	**0.499**	**58.065**	12.532	0.1839

Table 3
Results from the online studies with experts (NewsImages participants) and non-experts (recruited on Prolific). Image fit is assessed on a 5-point Likert scale, ranging from 1 (very poor fit) to 5 (very good fit). Best results are highlighted in bold.

Participants	Flux.2 Klein	Qwen-Image-Edit-2509	FireRed-Image-Edit-1.1	Editorial Baseline
Experts	3.271	**3.296**	3.258	3.021
Non-Experts	2.619	**2.807**	2.752	2.698
Average	2.990	**3.085**	3.044	2.880

3.2. Online Evaluation

Table 3 shows the results from the online user studies. We see that experts consistently rate image fit higher than non-experts. Qwen-Image-Edit-2509 is the best-performing approach across both participant pools. And almost all restoration workflows outperform the editorial baseline. The only exception to this is Flux.2 Klein. Non-experts did not perceive restored images with intentional flaws and visual artifacts as a better fit than the original images.

Overall, the results of the online studies do contrast with the numbers of the offline evaluation. The perceptual quality metrics favor Flux.2 Klein and FireRed-Image-Edit-1.1. The online evaluation, however, shows Qwen-Image-Edit-2509 as the winner across both studies. Offline perceptual metrics provide useful model-level signals. But they cannot fully capture the perceived fit between articles and images. To us, this signals that further research is needed to develop more effective offline evaluation regimes that better align with human perception.

4. Conclusion

In this work, we explored how restoring monochromatic news images with image-to-image models impacts perceived image fit. Our results show that both experts and non-experts find restored images more fitting than the originals. This indicates to us that image-to-image models are a viable option for restoring news visuals. The models allow the (re-)use of older images, e.g., to provide historical context for recent stories or to augment existing image collections.

There are, however, several limiting factors to keep in mind. The use of generative restoration in news contexts should be guided by clear disclosure and editorial safeguards[11] and consider the broader societal implications when used for providing news recommendations [11]. In our testing, we found that restoration is not always faithful and that models may swap people's gender or ethnicity, use the wrong alphabet for letters, and exhibit high uncertainty when using color. Therefore, human oversight and evaluation feedback remain necessary to assess whether restored images are considered trustworthy and editorially appropriate.

Acknowledgments

We thank Bram Bakker for providing the dataset of historical news articles. Qi Bi, Bruno N. Sotic, and Jaap Kamps are partly funded by the Netherlands Organization for Scientific Research (NWO NWA #1518.22.105). Jaap Kamps is further supported by the University of Amsterdam (AI4FinTech program) and ICAI (AI for Open Government Lab).

Declaration on Generative AI

The authors used GPT-5.5 and Grammarly for spelling checks. The authors have reviewed and edited the content as needed. They take full responsibility for the publication's content.

References

[1] L. Heitz, B. N. Sotic, A. A. Katamjani, Q. Bi, B. Bakker, L. Rossetto, J. Kamps, Newsimages in mediaeval 2026 - automated image recommendations with retrieval and generation techniques for news articles thumbnails, in: Working Notes Proceedings of the MediaEval 2026 Workshop, 2026.

[2] L. Heitz, L. Rossetto, B. Kille, A. Lommatzsch, M. Elahi, D.-T. Dang-Nguyen, Newsimages in mediaeval 2025 – comparing image retrieval and generation for news articles, in: Working Notes Proceedings of the MediaEval 2025 Workshop, 2025.

[3] L. Heitz, A. Bernstein, L. Rossetto, An empirical exploration of perceived similarity between news article texts and images, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[4] L. Heitz, Y. K. Chan, H. Li, K. Zeng, A. Bernstein, L. Rossetto, Prompt-based alignment of headlines and images using openclip, in: Working Notes Proceedings of the MediaEval 2023 Workshop, 2024.

[5] B. F. Labs, FLUX.2: Frontier Visual Intelligence, 2025.

[6] C. Wu, J. Li, J. Zhou, J. Lin, K. Gao, K. Yan, S.-m. Yin, S. Bai, X. Xu, Y. Chen, et al., Qwen-image technical report, arXiv preprint arXiv:2508.02324 (2025).

[7] S. I. Team, C. Qiao, C. Hui, C. Li, C. Wang, D. Song, J. Zhang, J. Li, Q. Xiang, R. Wang, et al., Firered-image-edit-1.0 technical report, arXiv preprint arXiv:2602.13344 (2026).

[8] S. Kastryulin, J. Zakirov, D. Prokopenko, D. V. Dylov, Pytorch image quality: Metrics for image quality assessment, 2022.

[9] Y. Kirstain, A. Polyak, U. Singer, S. Matiana, J. Penna, O. Levy, Pick-a-pic: An open dataset of user preferences for text-to-image generation, Advances in neural information processing systems 36 (2023) 36652–36663.

[10] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, Clipscore: A reference-free evaluation metric for image captioning, in: Proceedings of the 2021 conference on empirical methods in natural language processing, 2021, pp. 7514–7528.

[11] L. Heitz, O. Inel, S. Vrijenhoek, Recommendations for the recommenders: Reflections on prioritizing diversity in the recsys challenge, in: Proceedings of the Recommender Systems Challenge 2024, 2024, pp. 22–26.

[11] For details on the guidelines considered for the subtasks, we refer to the AI-CODE Deliverable D3.1 on user-centered requirements definition (AI-CODE Consortium, 2025).