Illustration Baseline for the NewsImages 2026 Challenge

Bruno N. Sotic[1], Jaap Kamps[1]

[1]University of Amsterdam, Amsterdam, the Netherlands

Abstract

We contribute an organizer baseline for the illustration direction of the NewsImages 2026 Challenge, complementing the photorealistic restoration baseline in our companion paper. We generate non-photorealistic visuals for the test set using three open-weight diffusion pipelines run at the default configuration: Flux.2 Klein and Qwen-Image-Edit-2509 in an image-to-image setting, taking the editorial original as a stylization input, and Z-Image-Turbo in a text-to-image setting, generating from the article title. We release all configurations, prompts, and generated images, and report ratings of perceived article image fit from an online user study with challenge participants and Prolific raters as a reference point for submissions to the 2026 challenge.

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 automatically matching news articles with appropriate visual accompaniments. The 2026 iteration introduces a mix of contemporary articles and scans of 19th- and 20th-century newspapers, and explicitly compares retrieval- and generation-based pipelines. Within this setup, the question of *how* a generated image relates to the article fit accompanies becomes increasingly central: not only whether the image fits the text, but also what kind of image it is, and what visual mode it operates in.

In a companion Quest for Insight [3], we examine *photorealistic restoration* of degraded historical scans, where diffusion-based editing models are used to enhance low-resolution monochromatic originals toward the appearance of contemporary photographs. That direction explicitly engages the tension between visual quality and authenticity: a restored photograph still presents itself as a photograph, and editorial use of such restorations requires careful disclosure and safeguards.[2] The present paper takes the opposite direction. Rather than restoring images toward photorealism, we focus on *illustration*: generating visuals that are explicitly non-photorealistic and that signal, by their visual form, that they are not photographs of an actual event.

This direction is motivated by two considerations. First, editorial guidelines around generative AI in news contexts increasingly require that AI-generated visuals be disclosed and visually distinguishable from photojournalistic imagery [4]. Stylized, illustrative imagery sidesteps the central concern of restored or photorealistic AI imagery: the risk that readers mistake a generated image for an evidentiary photograph. An illustration does not claim to depict the

MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, Netherlands and Online

*Corresponding author.

nadalic.sotic@uva.nl (B. N. Sotic); kamps@uva.nl (J. Kamps)

0009-0003-2122-2235 (B. N. Sotic); 0000-0002-6614-0087 (J. Kamps)

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

[2]See AI-CODE Deliverable D3.1 on user-centered requirements for editorial use of generative AI.

event; it visually communicates the theme. Second, prior NewsImages work has shown that AI-generated images can be perceived as a better fit than the editorially-assigned original [5], and that prompt-aligned generation can outperform retrieval over large image collections [6]. These findings together suggest that generation has a legitimate role in news image selection *provided* the visual mode is appropriate to the editorial context. Illustration is one such mode.

We frame this contribution as a *baseline* for the illustration direction of the 2026 challenge. As task organizers, our goal here is not to propose a novel or optimised stylization pipeline, but to provide a transparent, easily reproducible reference point against which participant submissions can be compared. We deliberately keep our approach minimal: we use three open-weight diffusion-based pipelines off the shelf, with simple prompts and default parameters, and we make all configurations, prompts, and generated images publicly available. The illustrations produced here should therefore be read as an organizer baseline rather than a stylistic benchmark; we expect participants' submissions to improve on these results in multiple dimensions.

We generate illustrative visuals using three complementary pipelines and evaluate them on the same article set and user study framework as the restoration companion paper. Details of the generation pipelines and experimental setup are provided in the following section.

2. Method

We generate illustrations with three open weight diffusion pipelines and contrast them with the editorial originals as a baseline (outputs are visible in the Appendix **??**). Two of the pipelines operate in an image to image setting, taking the original article image as a stylization input. The third operates in a text to image setting, generating an illustration directly from the article title. All configurations, prompts, generation parameters, and final outputs are released through our GitHub repository.[3]

The three pipelines are `FLUX.2-klein-base-9B`, `Qwen-Image-Edit-2509` together with a Lightning step distillation LoRA, and `Z-Image-Turbo`. Each is run at default or near default configuration; we have not tuned prompts or parameters per article.

Flux.2 Klein. We use the undistilled 9B base configuration of Flux.2 Klein[4] through the diffusers `Flux2KleinPipeline`, loaded in BF16. Inputs are the original article images, first scaled to approximately one megapixel with side lengths rounded to a multiple of sixteen pixels and original aspect ratio preserved. We use the prompt *"transform this photograph into an editorial newspaper political cartoon, hand drawn ink linework, exaggerated facial features, bold black outlines, flat color blocks, simplified shapes, halftone shading, no photorealism, 2D illustration"*, with twenty inference steps and a guidance scale of 5.0. Even in BF16, the 9B transformer combined with its text encoder exceeds the available 40 GB of GPU memory, so we enable sequential CPU offloading.

Qwen-Image-Edit-2509. We use `Qwen-Image-Edit-2509`[5] through the `QwenImageEditPlusPipeline` from diffusers, combined with the 4 step Lightning step distillation LoRA from `lightx2v/Qwen-Image-Lightning`.[6] The scheduler is a `FlowMatchEulerDiscreteScheduler` configured with dynamic shifting and exponential time shifting. Inputs are scaled as for Flux.2 Klein. We use the prompt *"turn this into a*

[3] https://github.com/Informfully/Challenges/tree/main/newsimages26/workflows/NewsImages
[4] https://huggingface.co/black-forest-labs/FLUX.2-klein-base-9B
[5] https://huggingface.co/Qwen/Qwen-Image-Edit-2509
[6] https://huggingface.co/lightx2v/Qwen-Image-Lightning

cartoon illustration", with four inference steps and a `true_cfg_scale` of 1.0 (no classifier free guidance, matching the Lightning LoRA's training conditions). As with Flux.2 Klein, we enable sequential CPU offloading.

Z-Image-Turbo. Unlike the previous two pipelines, Z-Image-Turbo[7] is used in a text to image setting: it generates an illustration directly from the article title rather than stylizing the editorial original. We use `Tongyi-MAI/Z-Image-Turbo` through the diffusers `ZImagePipeline` in BF16. For each article we construct a prompt of the form *"Editorial newspaper political cartoon illustration depicting: ⟨title⟩. Hand drawn ink linework, exaggerated facial features, bold black outlines, flat color blocks, simplified shapes, halftone shading, no photorealism, 2D illustration,"* where ⟨title⟩ is the article title from the test set metadata. Generation uses four inference steps at 1024×1024 with classifier free guidance disabled (`guidance_scale` $= 0.0$, equivalent to `cfg` $= 1.0$ in ComfyUI). All three pipelines produce outputs which we resize to the NewsImages thumbnail format of 460×260 pixels using Lanczos resampling and save as PNG following the naming convention ⟨`article_id`⟩`_Sotic_`⟨`approach`⟩`.png`, with a fixed random seed of 42 throughout. The Flux.2 Klein and Qwen-Image-Edit-2509 pipelines are applied to the same article subset used in the photorealistic restoration companion paper [7]; Z-Image-Turbo is additionally run over the full set of test articles, since its text to image setting does not require a per article input image.

3. Results and Analysis

We evaluate the three caricature pipelines through an online user study that mirrors the format used in our companion restoration paper [3]. Following the same protocol, perceived article–image fit is rated on a 5-point Likert scale ranging from 1 (very poor fit) to 5 (very good fit), across two participant pools: experts (NewsImages 2026 challenge participants, $N = 16$), and non-experts recruited on Prolific ($N = 30$). For each of 35 stimulus articles, each rater saw the editorial original alongside the three caricature variants and assigned a fit score to each. Expert ratings are reported with self-ratings excluded, so that no participant's score for their own submission is counted toward their group's average.

Table 1 summarizes the results. Two observations are immediate. First, experts rate every variant higher than Prolific raters do. This pattern is consistent with the restoration companion paper and likely reflects familiarity with the task and a more lenient calibration toward non-photorealistic outputs by participants who have themselves engaged with the generation problem. Second, and contrary to prior NewsImages findings on photorealistic generation [5] and to the restoration results of the companion paper, *all three caricature pipelines score below the editorial baseline across both participant pools.* The editorial original receives the highest fit rating in every column.

Among the caricature pipelines themselves, the ordering is consistent across both pools: Z-Image-Turbo is rated highest, Qwen-Image-Edit-2509 second, and Flux.2 Klein lowest. Several factors plausibly contribute to this ordering. Z-Image-Turbo generates from the article title rather than stylizing the original image, so its outputs are unconstrained by the composition, framing, or content of the editorial photograph; this gives it freedom to produce a visual that addresses the article's topic directly. Qwen-Image-Edit-2509 produces colored illustrations that retain the original's composition, landing in a middle ground. Flux.2 Klein is configured to produce explicitly black-and-white, hand-drawn-style outputs with intentional linework and

[7] https://huggingface.co/Tongyi-MAI/Z-Image-Turbo

Table 1
Online study results for the three caricature pipelines. Image fit is assessed on a 5-point Likert scale, from 1 (very poor fit) to 5 (very good fit). Experts are NewsImages 2026 challenge participants; the expert average excludes self-ratings. Best results among the caricature pipelines are highlighted in bold; the editorial baseline is reported for reference.

Participants	Flux.2 Klein	Qwen-Image-Edit-2509	Z-Image-Turbo
Experts ($N = 16$, no self-ratings)	2.162	2.640	**2.847**
Non-Experts (Prolific, $N = 30$)	1.755	1.973	**2.232**
Average	1.972	2.330	**2.563**
Editorial Baseline (Original Images)	Experts: 3.021	Prolific: 2.698	Avg: 2.880

halftone shading; these stylistic choices, while consistent with the cartoon aesthetic, may be read by raters as a poorer fit for contemporary news content, in which colored imagery is the dominant convention.

We caution against reading these results as a verdict on illustrative generation as a category. The pipelines reported here are organizer baselines, run at default or near-default configurations with a single prompt template per model and no per-article tuning. They are intended as a transparent reference point against which participant submissions can be compared, not as a stylistic benchmark. Several factors plausibly limit the fit scores observed here. The fit scale itself, which asks raters to evaluate how well an image accompanies an article, may implicitly favor photographic conventions; an illustration that captures the article's theme well but departs visually from photojournalistic convention is difficult to score on a single dimension. Likewise, the visual mode signaled by a caricature (explicitly non-photographic, hand-drawn, exaggerated) may be read by raters as inappropriate for serious news content even when the image's thematic match is strong.

4. Conclusion

In this work we documented three open-weight diffusion pipelines for generating illustrative, non-photorealistic visuals from news article content, contributed as an organizer baseline for the illustration direction of the NewsImages 2026 challenge. We described the configurations and prompts for Flux.2 Klein, Qwen-Image-Edit-2509, and Z-Image-Turbo, and evaluated perceived article–image fit on a 5-point scale through an online user study with 16 challenge participants and 30 Prolific raters across 35 articles. The pipelines, prompts, and all generated images are released through our GitHub repository as a reference point for future participant submissions.

The baselines reported here are intentionally simple (off-the-shelf pipelines run with a single prompt template per model and no per-article tuning) and we expect participant submissions to improve on them. We also note that perceived article–image fit, expressed as a single rating, is a limited instrument for evaluating non-photorealistic imagery: it does not separate thematic match from visual mode or from the indexical expectations readers carry into news contexts. We pursue a multi-dimensional treatment of this evaluation question, drawing on prior frameworks for journalistic image access and indexicality in mechanically reproduced, in separate companion work [8].

Acknowledgments

We thank Bram Bakker for providing the dataset of historical news articles, and Qi Bi and Lucien

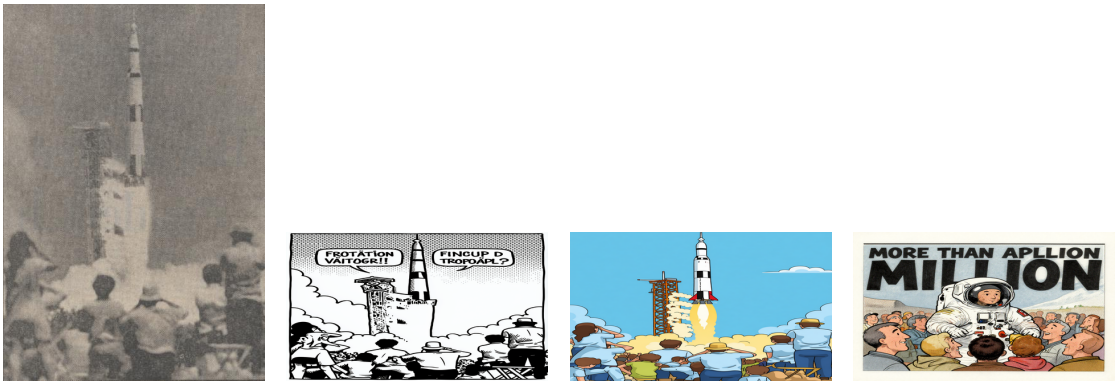
Heitz for providing the generation pipelines. Bruno N. Sotic, and Jaap Kamps are partly funded by the Netherlands Organization for Scientific Research (NWO NWA #1518.22.105). Jaap Kamps is further supported by the University of Amsterdam (AI4FinTech program) and ICAI (AI for Open Government Lab).

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.

A. Appendix

Figure 1 shows the editorial original alongside illustration outputs from the three pipelines for one article in the test set.



(a) Editorial original (b) Flux.2 Klein (c) Qwen-Image-Edit-2509 (d) Z-Image-Turbo

Figure 1: Example outputs from the three illustration pipelines for one article from the NewsImages 2026 test set, shown alongside the editorial original.

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, B. N. Sotic, Q. Bi, J. Kamps, Assessing the viability of image-to-image models for restoring historical photographs of news articles, in: Proceedings of the MediaEval'26: Multimedia Evaluation Workshop, Amsterdam, Netherlands, 2026. To appear.

[4] T. J. Thomson, R. J. Thomas, P. Matich, Generative visual ai in news organizations: Challenges, opportunities, perceptions, and policies, Digital Journalism (2024). doi:10.1080/21670811.2024.2331769.

[5] 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.

[6] 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.

[7] L. Heitz, B. N. Sotic, Q. Bi, J. Kamps, Assessing the viability of image-to-image models for restoring historical photographs of news articles, in: Working Notes Proceedings of the MediaEval 2026 Workshop, 2026.
[8] B. N. Sotic, J. Kamps, Operationalising journalistic and indexical dimensions of ai imagery, in: Proceedings of the MediaEval'26: Multimedia Evaluation Workshop, Amsterdam, Netherlands, 2026. To appear.