Diffusion-Based Approaches for NewsImage Generation: A Comparative Study of SDXL Variants

Xiaomeng Wang[1,*,†], Bram Bakker[1]

[1]Radboud University, Netherlands

Abstract

The MediaEval 2025 NewsImage Generation Challenge focuses on generating suitable thumbnails for news articles. We propose three SDXL-based methods: SDXL, which directly generates images from English news titles; SDXLREF, which enhances visual quality via latent-space refinement; and SDXLNEG, which incorporates handcrafted negative prompts to reduce common artifacts such as distorted faces. In a crowdsourced evaluation of image relevance and overall fit, our methods outperform the baseline on the *LARGE* subset by an average of 0.0414, but underperform on the *SMALL* subset by 0.0670.

1. Introduction

Visuals play a vital role in online news, significantly boosting user engagement and click-through rates. Prior studies show that AI-generated or retrieved images are often preferred over editorially selected ones [1], suggesting automated approaches may better align with audience preferences. Recent advances in generative models—such as Stable Diffusion XL (SDXL, [2]), FLUX[3], and GPT-4o [4]—enable automatic thumbnail generation from textual content. The MediaEval 2025 NewsImage task evaluates such retrieval–generation pipelines for this purpose [5].

This paper, from the Das-RU team, focuses on the generation subtask and proposes three SDXL-based methods. We adopt SDXL as the backbone due to its ability to generate high-resolution, semantically aligned, and visually coherent images. Its open-source nature and support also make it a suitable foundation for experimentation. Additionally, its two-stage base–refiner design enhances visual fidelity and structural consistency in generating human faces [2]. It is especially beneficial for news-related prompts involving human subjects. Built on SDXL, we design three pipelines: (a) **SDXL**: A baseline that generates images directly from news titles; (b) **SDXLREF**: A two-stage version with latent-space refinement for better clarity and detail; (c) **SDXLNEG**: A prompt-engineered variant using handcrafted negative prompts (e.g., blurry, bad faces) to reduce artifacts.

In the fully automated *LARGE* subset, our models outperform the baseline by an average of 0.0414. However, on the *SMALL* subset, they underperform by 0.0670. These results highlight the potential of generation-based methods at scale, while also raising broader questions about the role of AI-generated visuals in journalism—issues we briefly reflect on in this paper.

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online

*Corresponding author.

†These authors contributed equally.

xiaomeng.wang@ru.nl (X. Wang); bram.bakker2@ru.nl (B. Bakker)

2. Approaches

2.1. SDXL: Base Generation Method

Our baseline method SDXL employs the Stable Diffusion XL pipeline to generate images from news titles directly. Each prompt is the English news title provided in the dataset. This serves as our reference implementation for subsequent improvements.

2.2. SDXLREF: Latent Representation Refinement

While SDXL produces visually diverse outputs, we found frequent artifacts and inconsistencies (e.g., blurred regions, lack of facial detail). To mitigate this, we apply a latent-space refinement module—leveraging the native SDXL Refiner model—to enhance the output image from the first-stage generation. This two-stage pipeline is denoted SDXLREF.

2.3. SDXLNEG: Negative Prompt Engineering

To further improve visual fidelity, especially for human-centric thumbnails, we introduce carefully crafted negative prompts. These are passed to the second text encoder to steer the generation away from common artifacts (e.g., blurry, deformed, ugly, low-res, bad faces and hands, text, missing fingers). We denote this approach SDXLNEG.

3. Experiments

3.1. Experimental Settings

We choose the `stable-diffusion-xl-base-1.0`[1] model released by StabilityAI on Hugging Face as the base model for all SDXL-based methods. All images are generated using a custom multi-GPU inference script built with PyTorch and the Hugging Face `diffusers` library. To ensure reproducibility, the random seed is fixed to 12345. All experiments are conducted on an HPC cluster equipped with NVIDIA A100 GPUs. Our code is publicly available [2].

Specifically, for the SDXL method, the generation configurations are as follows: image resolution of 1024×1024 pixels, 40 inference steps, a classifier-free guidance scale of 7.5, and mixed precision (`fp16`). All outputs are saved in PNG format. For the SDXLREF method, the second-stage refinement model is `stable-diffusion-xl-refiner-1.0` model[3]. Both the base and refiner pipelines are run in `fp16` precision with attention slicing enabled to manage GPU memory. The total number of inference steps remains 40. The denoising process is divided into two stages: the base model handles the first 80% (`denoising_end=0.8`) to produce latent representations, which are then passed to the refiner for the final 20% (`denoising_start=0.8`) to improve visual quality. For the SDXLNEG method, we extend the base SDXL pipeline by incorporating a handcrafted negative prompt: *"blurry, deformed, ugly, lowres, bad faces and hands, text, missing fingers"*. The model and generation parameters remain consistent with the SDXL configuration.

[1] https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0
[2] https://github.com/XiaomengWang-AI/MediaEval2025_NewsImage_Das-RU_GEN
[3] https://huggingface.co/stabilityai/stable-diffusion-xl-refiner-1.0

3.2. Evaluation results

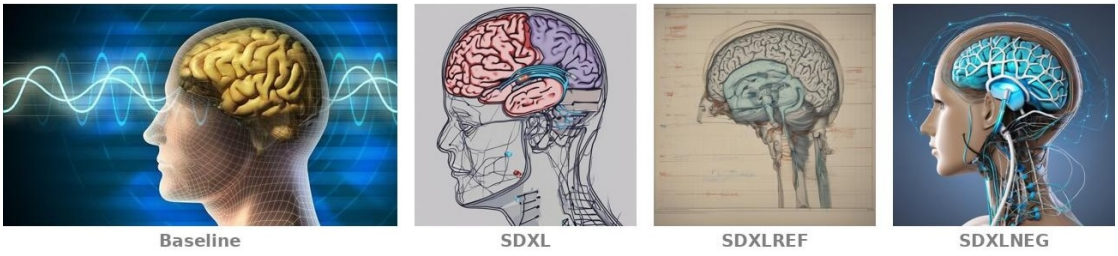
As shown in Table 1, all three methods outperform the baseline on the *LARGE* subset, with SDXL achieving the highest gain (+0.0744). In contrast, all methods underperform on the *SMALL* subset. Averaged across both subsets, SDXL and SDXLREF show net improvements over the baseline, while SDXLNEG performs worse overall.

Qualitatively, the topic diversity of *LARGE* subset enables generative models to produce more contextually appropriate and varied images, highlighting their effectiveness in large-scale, fully automated settings. Conversely, the *SMALL* subset demands greater semantic precision and visual specificity, where models occasionally generate generic or misaligned outputs, suggesting room for improvement in controlled editorial workflows.

Among the methods, SDXLREF effectively reduces noise and enhances structural consistency—especially for human faces and body parts—but sometimes sacrifices semantic alignment, as reflected in its scores. SDXLNEG, designed to suppress visual artifacts via negative prompts, offers only marginal visual improvements and underperforms in semantic relevance, falling short of its intended effect.

Table 1
Average image-fit ratings of baseline and proposed methods on the SMALL and LARGE subsets. Scores were collected through a crowdsourced online evaluation using a 5-point Likert scale, where 1 indicates poor fit and 5 indicates excellent fit.

Subsets	Baseline	SDXL	SDXLREF	SDXLNEG
SMALL	3.0412	2.9710	2.9838	2.9677
LARGE	2.9563	3.0307	3.0242	2.9379
Average	2.9988	3.0008	3.0040	2.9528



(a) AI-Guided Brain Stimulation Improves Memory in Patients With Epilepsy, TBI

(b) Idaho college killings: Prosecutors dispute DNA could have been planted at scene

Figure 1: Examples generated from three methods and baseline. The news title is shown as the subcaption of each subfigure.

4. Discussion and Outlook

Our paper suggests that diffusion-based generative models, particularly Stable Diffusion XL, are promising tools for large-scale news image generation. However, enhancements such as refinement stages and prompt engineering introduce nuanced trade-offs between image quality, semantic relevance, and model behavior. In the following sections, we reveal several insights that go beyond quantitative performance.

Generated vs. Archival Images: Which is Better?

The fitness of generated images compared to archival images depends largely on the nature of the news content. In the LARGE subtask, which involves automatic generation across a broad range of abstract and diverse topics, our generative methods often outperform the baseline.

However, as shown in the Figure 1b, for news involving identifiable individuals (e.g., politicians, celebrities) or concrete real-world events, archival images remain more appropriate due to their factual accuracy. Generative models, while creative, may introduce hallucinated or generic elements that risk misrepresenting the news. Thus, generated images are better suited for abstract or opinion-driven stories (as shown in the Figure 1a), while archival images are preferable for fact-based reporting.

Key Parameters for High-Quality and Relevant Generation

Our experiments show that the most influential factor in generating relevant and high-quality images is the specificity and clarity of the input prompt. Titles that include proper nouns, concrete actions, or clear emotional cues tend to produce images that are both visually appealing and semantically aligned.

Additional insights include:

- **Prompt length and structure** matter: overly short or vague prompts often lead to generic and less relevant outputs.
- **Latent refinement**, as applied in SDXLREF, enhances visual consistency and realism, especially for human-centric prompts, though it may slightly reduce semantic alignment.
- **Negative prompts** show some effectiveness in suppressing visual artifacts (e.g., facial distortions, missing limbs), though not as strongly as anticipated.

Effectively balancing prompt richness with model constraints is essential for achieving both visual fidelity and semantic relevance.

Which News Categories Work Best for Generation or Retrieval?

Our analysis indicates that the suitability of generation or retrieval varies across news categories:

- **Best suited for generation:** Technology, environment, lifestyle, and opinion pieces benefit from generative models, especially when the article's message is conceptual or symbolic. In such cases, the model can create abstract or stylized imagery that complements the news title.
- **Best suited for retrieval:** Hard news topics—such as politics, public safety, crime, and sports—demand factual grounding and visual accuracy. Here, real photographs offer credibility and specificity that generative models cannot yet reliably replicate.

These findings suggest that a hybrid system, which dynamically chooses between generation and retrieval based on topic classification or metadata, could offer an optimal solution for real-world deployment.

5. Conclusion

This paper presents Stable Diffusion XL–based pipelines for the NewsImage generation subtask at MediaEval 2025. Our results demonstrate that diffusion models can generate high-quality, semantically relevant thumbnails, particularly for abstract or conceptual news. While generative methods offer flexibility, archival retrieval remains preferable for fact-based reporting. We conclude that a hybrid strategy—selecting between generation and retrieval by topic—offers the most promising direction.

Declaration on Generative AI

During the preparation of this work, the authors used GPT-4o and Grammarly in order to: Grammar and spelling check. After using these tools, the authors reviewed and edited the content as needed and takes full responsibility for the publication's content.

References

[1] 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.

[2] D. Podell, Z. English, K. Lacey, A. Blattmann, T. Dockhorn, J. Müller, J. Penna, R. Rombach, SDXL: Improving latent diffusion models for high-resolution image synthesis, in: Proceedings of the Twelfth International Conference on Learning Representations, 2024.

[3] S. Batifol, A. Blattmann, F. Boesel, S. Consul, C. Diagne, T. Dockhorn, J. English, Z. English, P. Esser, S. Kulal, et al., Flux. 1 kontext: Flow matching for in-context image generation and editing in latent space, arXiv e-prints (2025).

[4] S. Chen, J. Bai, Z. Zhao, T. Ye, Q. Shi, D. Zhou, W. Chai, X. Lin, J. Wu, C. Tang, et al., An empirical study of gpt-4o image generation capabilities, arXiv preprint arXiv:2504.05979 (2025).

[5] 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.