

Less is More: A Dual-Filtering Perspective on Prompt Design for News Thumbnail Generation

Thanh-Khoi Nguyen^{1,2,†}, Hai-Dang Nguyen^{1,2} and Minh-Triet Tran^{1,2,*,†}

¹Faculty of Information Technology & Software Engineering Lab, University of Science, Ho Chi Minh city, Vietnam

²Vietnam National University, Ho Chi Minh city, Vietnam

Abstract

Visual elements play a pivotal role in modern journalism, functioning as powerful vehicles for conveying information, evoking emotions, and fostering audience engagement. In this paper, our team, SELab-HCMUS, propose our approach to multimedia content generation, focusing on the task of producing the most fit and relevant thumbnail images for news articles. Our framework leverages Large Language Models (LLMs) with a two-stage refinement process to generate optimized prompts, which are subsequently refined and fed into a Text-to-Image (T2I) model. To enhance diversity in the generated outputs, we augment each prompt with varied stylistic elements and alternative textual inputs, allowing the model to generate different visual thumbnails of the same article.

1. Introduction

Thumbnail images function as the primary visual anchor in digital news, shaping a reader's initial perception and guiding their engagement with the content. The NewsImages2025 [1] benchmark provides a structured venue for exploring this domain, challenging researchers to develop systems for both image retrieval and generation based on fitness and relevance. Unlike literal image captioning, this task is particularly difficult due to the abstract, often thematic or metaphorical, relationship between a news text and its image, demanding a nuanced semantic understanding of what constitutes a proper fit while avoiding the introduction of non-factual elements. Furthermore, an empirical study [2] demonstrated that the benchmark's core assumption of a single ground-truth image fails to capture the inherent ambiguity of perceived relevance, as human evaluators often find multiple, diverse images to be equally suitable. In response to this methodological gap, we propose a generative framework designed to embrace this diversity by producing multiple distinct runs, each exploring varied stylistic and methodological approaches. We utilize Llama3-8B Instruct¹ to distill the semantic core of each article, capitalizing on the proven efficacy of LLMs in news summarization [3]. The raw LLM output then undergoes a two-stage refinement process to ensure both fitness and relevance: Coverage Filter and Conciseness Filter. To produce a diverse set of candidates, we systematically augment these refined prompts with distinct stylistic directives and experiment with different textual inputs. Counter-intuitively, our empirical results reveal that prompts from the article's title often perform slightly better than those from refined LLM-generated summaries.


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


*Corresponding author.

[†]These authors contributed equally.

✉ 23120009@student.hcmus.edu.vn (T. Nguyen); tmtriet@fit.hcmus.edu.vn (M. Tran)

ORCID 0009-0003-5879-451X (T. Nguyen); 0000-0003-0888-8908 (H. Nguyen); 0000-0003-3046-3041 (M. Tran)

 © 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Source code available at: https://github.com/NTKHarry/MediaEval2025_NewsImage.git

¹<https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct>

2. Related Work

The rapid progress of T2I generation, together with its growing public accessibility, has shifted AI-generated visuals from being a specialized experimental technology into a mainstream tool for creative media production. Early foundational methods were based on Generative Adversarial Networks (GANs) [4, 5, 6], which rely on a dual-network paradigm: a generator produces images conditioned on textual input, while a discriminator learns to differentiate between authentic and synthetic samples.

In recent years, Diffusion Models (DMs) [7, 8, 9] have become the dominant paradigm for text-to-image synthesis, achieving superior visual quality and diversity compared to GAN-based approaches. An important advancement within this family is the Latent Diffusion Model (LDM) [10, 11, 12], which enhances efficiency by performing the diffusion process in a compressed latent representation rather than the high-dimensional pixel domain, thereby reducing computational demands while preserving high-fidelity image generation. Despite their impressive generative capabilities, current text-to-image models exhibit significant limitations when applied to the rigorous demands of news content generation.

3. Approach

3.1. Overview

Our approach employs a single LLM with a two-phase prompt filtering mechanism to ensure alignment and faithfulness to the given article. We adopt Infinity 8B [13] for two main reasons. First, it achieves strong performance on human preference benchmarks such as HPS [14] and ImageReward [15]. Second, Infinity’s bitwise token prediction and scalable autoregressive architecture enable it to process complex prompts with greater consistency and fidelity. The overview of our framework is shown in Figure 1.

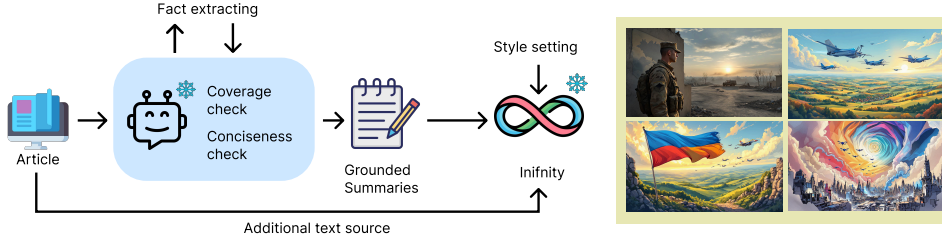


Figure 1: Overview of our pipeline. We first extract k key facts from the article, which serve as the backbone for summary generation. Each summary then undergoes a two-phase filtering process and is augmented with stylistic guidance to promote diversity in the resulting images. Our framework support generation for abstract, cartoon, realistic and modern art style.

3.2. Fact Extraction and Prompt Refinement

Since article titles typically provide only brief and condensed information, relying solely on the title as a generation prompt often omits key details from the article. As a result, the generated images may mislead readers by failing to accurately reflect the article’s content. Moreover, without access to the full article text, the LLM cannot properly align the image style with the article’s intended theme (e.g., adopting a serious tone for political news or a sarcastic tone for satirical pieces). To address this, we retrieve full web content to enrich inputs and generate faithful, stylistically consistent prompts.

After enriching each article with its full content, we employ Llama3-8B Instruct with a temperature setting of 0.7 to extract k key facts from each article, which serve as the backbone for summary generation. Here, k acts as a hyperparameter that can be adjusted in future experiments; for demonstration purposes, we set $k = 10$. The same model is then used to produce the initial summaries. To further ensure the faithfulness and conciseness of these summaries, we apply a two-phase filtering process:

Coverage Check. To prevent information loss, we enforce a coverage constraint that guarantees all extracted key facts are represented in the generated summaries. Specifically, let $\text{contains}(s, k)$ be a boolean function indicating whether a summary sentence s semantically includes fact k . The coverage requirement is satisfied when $\forall k_i \in K, \exists s_j \in S$ such that $\text{contains}(s_j, k_i)$. This ensures that the summary provides a comprehensive reflection of the article’s core content.

Conciseness Check. To eliminate redundancy and ungrounded statements, we require that each candidate summary sentence references at least one key fact. Sentences s_j that do not satisfy the condition $\exists k_i \in K$ such that $\text{contains}(s_j, k_i)$ are discarded. By filtering out such sentences, the process maintains a high signal-to-noise ratio, reduces semantic drift, and results in summaries that are both faithful and concise.

The refined summaries are then used as prompt for image generation. For diversity, we vary the image style to the prompt (e.g, cartoon, realistic, abstract and modern art style).

4. Results and Analysis

The primary evaluation metric for the task is image fit, assessed by human raters on a 5-point Likert scale. A higher score indicates a stronger alignment between the image’s content and the key aspects of the news article, without introducing non-factual elements. In total, we conduct six experimental runs, all based on the pipeline described in Section 3, but differing in the form of input provided and styling.

First 2 runs - Raw title: We simply use the articles’ title without any additional information source, with the image style of cartoon and realistic respectively.

Next 2 runs - Refined summaries: We employ our pipeline to retrieve the refined summaries, serving as input prompt for generation, with the same image style of cartoon and realistic.

Final 2 runs - LLM as Judge: LLMs can serve as assistants or evaluators [16], the final two runs are dedicated to leveraging LLMs to select the most suitable image from our pool of generated candidates. Specifically, we adopt the same pipeline described in Section 3, using both text sources (summaries and titles) each with four different styles (cartoon, modern, abstract, and modern art). Subsequently, we employ three multi-modal large language models (MLLMs): two act as candidates, each selecting the image style that best aligns with the article content, while the third serves as a judge to evaluate the candidates’ selections. Given that LLMs often struggle with direct hard-label assignments [17], we design a scoring mechanism instead. Each image is initialized with a fixed credit, and the MLLMs assign scores to the images within an article. The final score for an image is computed as a weighted average of the three models with respect to the article content, formulated as: $\text{score}_i = \frac{\alpha \cdot \text{candidate}_1 + \beta \cdot \text{candidate}_2 + \gamma \cdot \text{judge}}{\alpha + \beta + \gamma}$. Currently, we set $\alpha = 1, \beta = 1, \gamma = 2$ respectively.

As shown in Table 1, our best-performing approach (Run 5) achieved an average image fit score of 2.98 which is slightly higher than the baseline (2.96). Interestingly, models conditioned on refined summaries perform slightly worse than those using only the article title. This finding

Table 1

Quantitative results of our runs compared to the official baseline. Scores represent the average image fit evaluated by human raters on a 5-point Likert scale.

Run ID	Description	Styling	Average Rating
1	Raw Title	Cartoon	2.93
2	Raw Title	Realistic	2.85
3	Refined Summaries	Cartoon	2.82
4	Refined Summaries	Realistic	2.86
5	LLM as Judge (Title)	Mixed	2.98
6	LLM as Judge (Summaries)	Mixed	2.92
-	Baseline	-	2.96

contrasts with our initial expectation: since summaries are designed to capture the fine-grained details of the article, we anticipated that images generated from them would more faithfully represent the full content and thus yield better results. We argue that this phenomenon can be better understood by considering the intrinsic nature of the provided dataset. The article titles are not arbitrary strings of text; rather, they are carefully composed and curated by professional journalists. In the process of writing, editing, and publishing, titles are deliberately condensed into short, highly informative phrases that highlight the most engaging and newsworthy aspects of the article. As such, they act as precise semantic anchors, steering both the reader’s and the model’s initial interpretation toward the central theme of the content.

By contrast, the lower performance observed from summaries in our study demonstrates that summaries—although richer in detail—tend to capture a broad spectrum of fine-grained facts and contextual nuances. While this comprehensiveness is valuable for human readers, it can inadvertently introduce peripheral or less critical elements that dilute the core signal. For text-to-image generation, such additional details function as noise, leading to outputs that may diverge from the central narrative or fail to emphasize the key engagement points.

5. Discussion and Outlook

Our study provides empirical insights into the interplay between textual input and image generation quality in the context of news thumbnails. While our initial expectation was that summaries, by capturing fine-grained details, would offer richer guidance for image synthesis, our results reveal that titles consistently outperform summaries. This outcome highlights the unique role of journalistic titles as condensed, high-salience cues, crafted to capture reader attention and encapsulate the essence of an article in a minimal form. Moreover, the selection of prompts, the number of key facts and the stylistic treatment of images in relation to the article’s genre also influence how readers perceive the news, with visual styling in particular being strongly shaped by individual aesthetic preferences.

Declaration on Generative AI

During the preparation of this work, the authors used ChatGPT, in order to: Paraphrase and reword. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the publication’s content.

References

- [1] 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.
- [2] 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.
- [3] L. R. P. Houamegni, F. Gedikli, Evaluating the effectiveness of large language models in automated news article summarization, 2025. URL: <https://arxiv.org/abs/2502.17136>. arXiv: 2502.17136.
- [4] A. Brock, J. Donahue, K. Simonyan, Large scale GAN training for high fidelity natural image synthesis, 2019. URL: <https://arxiv.org/abs/1809.11096>. arXiv: 1809.11096.
- [5] I. J. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, 2014. URL: <https://arxiv.org/abs/1406.2661>. arXiv: 1406.2661.
- [6] T. Karras, S. Laine, T. Aila, A style-based generator architecture for generative adversarial networks, 2019. URL: <https://arxiv.org/abs/1812.04948>. arXiv: 1812.04948.
- [7] 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, 2023. URL: <https://arxiv.org/abs/2307.01952>. arXiv: 2307.01952.
- [8] J. Chen, J. Yu, C. Ge, L. Yao, E. Xie, Y. Wu, Z. Wang, J. Kwok, P. Luo, H. Lu, Z. Li, Pixart- α : Fast training of diffusion transformer for photorealistic text-to-image synthesis, 2023. URL: <https://arxiv.org/abs/2310.00426>. arXiv: 2310.00426.
- [9] J. Chen, C. Ge, E. Xie, Y. Wu, L. Yao, X. Ren, Z. Wang, P. Luo, H. Lu, Z. Li, Pixart- σ : Weak-to-strong training of diffusion transformer for 4k text-to-image generation, 2024. URL: <https://arxiv.org/abs/2403.04692>. arXiv: 2403.04692.
- [10] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, B. Ommer, High-resolution image synthesis with latent diffusion models, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2022, pp. 10684–10695.
- [11] X. Dai, J. Hou, C.-Y. Ma, S. Tsai, J. Wang, R. Wang, P. Zhang, S. Vandenhende, X. Wang, A. Dubey, M. Yu, A. Kadian, F. Radenovic, D. Mahajan, K. Li, Y. Zhao, V. Petrovic, M. K. Singh, S. Motwani, Y. Wen, Y. Song, R. Sumbaly, V. Ramanathan, Z. He, P. Vajda, D. Parikh, EMU: Enhancing image generation models using photogenic needles in a haystack, 2023. URL: <https://arxiv.org/abs/2309.15807>. arXiv: 2309.15807.
- [12] P. Esser, S. Kulal, A. Blattmann, R. Entezari, J. Müller, H. Saini, Y. Levi, D. Lorenz, A. Sauer, F. Boesel, D. Podell, T. Dockhorn, Z. English, K. Lacey, A. Goodwin, Y. Marek, R. Rombach, Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL: <https://arxiv.org/abs/2403.03206>. arXiv: 2403.03206.
- [13] J. Han, J. Liu, Y. Jiang, B. Yan, Y. Zhang, Z. Yuan, B. Peng, X. Liu, Infinity: Scaling bitwise autoregressive modeling for high-resolution image synthesis, 2025. URL: <https://arxiv.org/abs/2412.04431>. arXiv: 2412.04431.
- [14] X. Wu, K. Sun, F. Zhu, R. Zhao, H. Li, Human preference score: Better aligning text-to-image models with human preference, 2023. URL: <https://arxiv.org/abs/2303.14420>. arXiv: 2303.14420.
- [15] J. Xu, X. Liu, Y. Wu, Y. Tong, Q. Li, M. Ding, J. Tang, Y. Dong, ImageReward: Learning and evaluating human preferences for text-to-image generation, 2023. URL: <https://arxiv.org/abs/2304.05977>. arXiv: 2304.05977.
- [16] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, Y. Liu, Llm-as-judges: A comprehensive survey on llm-based evaluation methods, 2024. URL: <https://arxiv.org/abs/2412.05579>. arXiv: 2412.05579.
- [17] S. Mandal, X. Lin, R. Srikant, A theoretical analysis of soft-label vs hard-label training in neural networks, 2024. URL: <https://arxiv.org/abs/2412.09579>. arXiv: 2412.09579.