Generating Images from News Titles via Structured Prompt Generation

Runze Li[1,*], Hong Qu[2]

[1]Hong Kong University of Science and Technology, Hong Kong
[2]Hong Kong Polytechnic University, Hong Kong

Abstract

This paper describes our submission to the MediaEval 2026 NewsImages task. We address title-based news image generation as structured prompt generation: a Qwen2.5-3B-Instruct model takes a news title and, in a single inference pass, produces title understanding, a visual generation plan, and a final prompt for non-photorealistic editorial image generation. We compare a zero-shot prompt generator with a LoRA-adapted version trained on title-to-plan examples generated by a teacher model using paired training images as weak visual guidance. For each prompt-generation setting, we generate images with SDXL and FLUX.1-schnell, resulting in four submitted runs. The official survey results show that zero-shot FLUX achieved the highest score among our submissions and scored slightly above the original images.

1. Introduction

The MediaEval NewsImages task studies how news articles can be matched with suitable visual thumbnails. In the 2026 edition, systems may retrieve existing images, generate new images from article titles, or combine both approaches [1]. The task encourages non-photorealistic outputs that do not suggest an accurate depiction of real events.

We focus on the generation setting. Generating images from news titles is challenging because a title is a compact semantic summary, not a visual description. It may mention entities, topics, actions, or news angles, but often leaves open the concrete subjects, objects, setting, composition, and editorial style required by a text-to-image model. This creates a semantic-to-visual gap: before generating an image, the system must interpret the title and turn it into explicit visual guidance.

We therefore formulate title-based news image generation as structured prompt generation. Given a title, an instruction-tuned Qwen2.5-3B-Instruct model [2] produces title understanding, a visual generation plan, and a final prompt for non-photorealistic editorial image generation in a single inference pass. This tests whether a general instruction-tuned model can perform title-to-visual prompt generation without task-specific training.

We further ask whether paired training images can provide useful weak supervision for adapting the prompt generator. The NewsImages training set provides titles and associated images, but not human-written prompts for non-photorealistic image generation. Moreover, each paired image is only one possible editorial choice, not the unique target image. It can provide visible facts such as people, objects, settings, and actions, but these facts depend on the title for their interpretation. For example, a hand counting money could be related to

MediaEval'25: Multimedia Evaluation Workshop, October 25–26, 2025, Dublin, Ireland and Online
*Corresponding author.
†These authors contributed equally.
runzeli@ust.hk (R. Li); hong.qu@connect.polyu.hk (H. Qu)

inflation, taxation, salary, or expenses. We therefore use the paired image only as weak visual evidence: a vision-language model extracts visible facts from the image, and a larger teacher model combines these facts with the title to construct title-to-plan examples. A LoRA-adapted [3] Qwen2.5-3B model is then trained to generate structured prompts from the title alone.

Finally, we use the prompts produced by the zero-shot and LoRA-adapted settings to generate images with two fixed models, SDXL [1] and FLUX.1-schnell [2]. This gives four submitted runs: zero-shot SDXL, zero-shot FLUX, LoRA SDXL, and LoRA FLUX. Since the image generators are not fine-tuned, this setup allows us to compare the effect of prompt generation settings and image generation backends. We analyze the official survey ratings and prompt diagnostics to understand which parts of the pipeline are most effective for title-based news image generation.

2. Related Work

Prior work on news images has studied how textual news content can be connected to visual representations in both retrieval and generation settings. In the retrieval setting, several systems address the semantic-to-visual gap by enriching or reformulating the textual input before image matching. Heitz et al. [4] explored prompt-based alignment between headlines and images using OpenCLIP, including title preprocessing, tags, text rewriting, and entity extraction. Their results suggest that transforming headlines into better visual queries is non-trivial, since simple reformulations do not necessarily improve over using the raw headline. Our work is related to this direction because it also addresses the mismatch between news titles and visual representations, but we focus on text-to-image generation rather than image retrieval. For news image generation, Wang and Bakker [5] compared SDXL-based variants, including direct title-to-image generation and variants with refinement or negative prompting. Opal [6] investigates multimodal image generation for news illustration and shows that structured exploration of article tone, keywords, and style can help users create more usable news illustrations. These works suggest that news image generation can benefit from an intermediate step that makes visual intent explicit before image generation. Our work follows this direction, but focuses on automatic title-only structured prompt generation: an instruction-tuned language model converts each title into title understanding, a visual generation plan, and a final prompt for off-the-shelf text-to-image models.

3. Approach

We formulate the task as title-to-prompt-to-image generation. Using the news title directly as a text-to-image prompt is a natural baseline, since it tests whether the title alone provides sufficient visual guidance. However, in our preliminary experiments, direct title prompting often produced generic or weakly grounded images, including pseudo-text, logo-like graphics, misleading literal symbols, or otherwise unsuitable visualizations. We therefore do not include raw-title prompting in our submitted runs. Instead, we focus on a structured prompt-generation step to make the intended subject, context, visual cues, and style constraints more explicit before image generation.

Our prompt generator is based on Qwen2.5-3B-Instruct. Given a news title, it produces a structured output in one pass: a short title interpretation, a visual plan, and a final prompt for the image generator. The title interpretation identifies the likely event or issue, important

[1] https://huggingface.co/docs/diffusers/using-diffusers/sdxl
[2] https://huggingface.co/black-forest-labs/FLUX.1-schnell

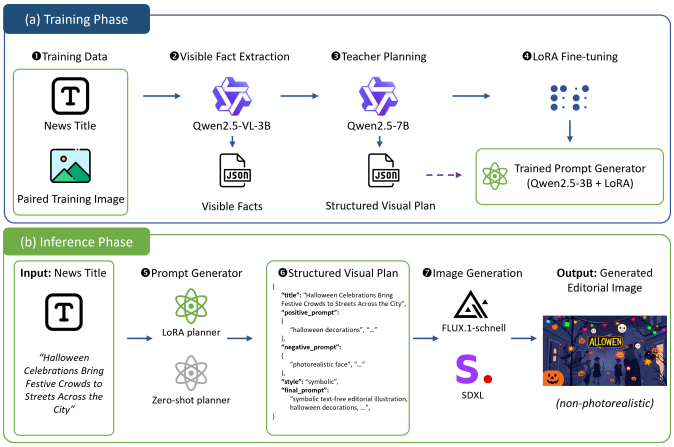

Figure 1: Overview of the structured prompt generation framework.

Table 1
Prompt diagnostics for the zero-shot and LoRA planners.

Planner	Words	Title cov.	Phrase cov.	Hygiene issues
Zero-shot	36.3	0.590	0.428	0.215
LoRA	19.0	0.224	0.078	0.001

entities, and their relations. The visual plan then turns this interpretation into concrete image guidance, including the likely subject, scene, visual style, and details to avoid, such as readable text, logos, or overly realistic faces. Finally, the model converts the plan into a concise prompt for non-photorealistic editorial image generation. Figure 1 shows the full pipeline.

We compare two prompt-generation settings. The first is zero-shot: Qwen2.5-3B-Instruct follows a fixed instruction and generates the structured prompt directly from the title. This tests whether a general instruction-tuned model can perform the title-to-prompt step without additional training. The second setting uses LoRA adaptation[3]. Since the training set contains titles and paired images, but no human-written prompts, we create training examples automatically. For each training article, a vision-language model extracts visible facts from the paired image, such as people, objects, settings, and actions. We use these facts only as weak visual guidance, because the paired image is just one possible editorial choice, not the unique correct image. A larger Qwen2.5-7B-Instruct [7, 8] teacher then combines the title and visible facts to write a structured visual plan and final prompt. These teacher outputs are used to fine-tune the Qwen2.5-3B prompt generator with LoRA.

The resulting prompts are passed to two image-generation backends, SDXL and FLUX.1-schnell. For SDXL, we use both the positive prompt and the negative constraints. For FLUX.1-schnell, we use only the positive prompt and prepend a fixed phrase to encourage a symbolic editorial style. This produces four submitted runs: zero-shot prompt result with SDXL (ZS-SDXL), zero-shot prompt result with FLUX (ZS-FLUX), LoRA prompt with SDXL (LoRA-SDXL), and LoRA prompt with FLUX (LoRA-FLUX).

4. Results and Analysis

We first analyze the prompts using deterministic lexical diagnostics. *Words* is the average prompt length. *Title coverage* is the fraction of title content words that also appear in the prompt.

Table 2
Human survey ratings for our submitted runs. Scores are the official overall averages, where Prolific ratings are counted as individual ratings.

Image source	Score
Original images	2.880
ZS-FLUX	**2.907**
LoRA-FLUX	2.826
ZS-SDXL	2.636
LoRA-SDXL	2.331

Phrase coverage is the fraction of title bi-grams preserved in the prompt, which provides a stricter measure of title grounding. *Hygiene issues* is the fraction of prompts containing obvious formatting artifacts such as placeholders, angle brackets, or underscores.

Table 1 shows a trade-off between prompt regularity and title grounding. The LoRA-version Qwen3B model produces shorter and cleaner prompts, and hygiene issues almost disappear. However, LoRA also preserves less title words and phrases. This suggests that LoRA regularizes the prompt format, but may over-compress the title and remove phrases that are important for news image generation.

For image-level evaluation, we report the official human survey scores provided by the organizers. Table 2 show that FLUX-based runs were rated higher than SDXL. The zero-shot FLUX run achieved the highest final score, slightly above the original-image baseline reported by the organizers. FLUX appears better suited to our structured editorial prompts than SDXL.

To inspect article-level patterns, we conduct a small post-hoc analysis of the surveyed items. The survey sheet contains 35 evaluated article blocks but only 34 unique article IDs, because one article appears twice. For article-level analysis, we average the duplicated entries before counting unique articles. After deduplication, the best generated image among our four submitted runs scores higher than the original image for 23 of 34 unique articles.

We further inspect the survey results by article category. We first attempted automatic category assignment using the zero-shot category prediction model from Informfully Recommenders [9], but the resulting labels were noisy for short and ambiguous headlines. We therefore manually assigned broad categories for qualitative interpretation. The inspection suggests that the main issue is not only the topic of the article, but the role expected from the image. In some stories, the image mainly serves as a visual interpretation of the title. For example, in the article related to "volunteer work," all four generated images receive higher ratings than the original image. In such cases, generated illustrations can be competitive because they do not need to document one specific real-world moment.

Disaster-related stories show a different requirement. In the survey subset, all three disaster articles receive higher ratings for the original image than for any generated image. These titles involve bushfire smoke, haze, and wildfire damage. In such cases, the image is expected not only to illustrate the topic, but also to provide documentary evidence of the actual event. This suggests that generated non-photorealistic images are less suitable for severe real-world events that require factual visual grounding.

Finally, LoRA adaptation improves prompt regularity but does not guarantee better image ratings. The prompt diagnostics show that LoRA produces shorter and cleaner prompts, but with substantially weaker title grounding. This helps explain why LoRA does not consistently outperform zero-shot planning: cleaner prompts may still lose concrete title-specific information that is important for news image generation.

5. Discussion and Outlook

Our results suggest three main findings. First, the image generator choice is important for this pipeline: both FLUX runs are rated higher than the SDXL runs, and zero-shot FLUX achieves the best score among our submissions. Second, LoRA improves prompt regularity but not image ratings. The prompt diagnostics show that LoRA produces shorter and cleaner prompts, but preserves fewer title words and phrases. Third, the post-hoc article analysis suggests that the suitability of generated images may depend on the expected role of the news image.

Future work should more systematically study what kind of news images are expected, for example, to provide visual interpretation or factual evidence. This could inform when generation, retrieval, or human editorial review is more appropriate.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-5.5 in order to: Improve writing style, Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) 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] Q. Team, Qwen2.5-vl, 2025. URL: https://qwenlm.github.io/blog/qwen2.5-vl/.

[3] E. J. Hu, yelong shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, W. Chen, LoRA: Low-rank adaptation of large language models, in: International Conference on Learning Representations, 2022. URL: https://openreview.net/forum?id=nZeVKeeFYf9.

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

[5] X. Wang, B. Bakker, Diffusion-based approaches for newsimage generation: A comparative study of sdxl variants (2025).

[6] V. Liu, H. Qiao, L. Chilton, Opal: Multimodal image generation for news illustration, in: Proceedings of the 35th Annual ACM Symposium on User Interface Software and Technology, UIST '22, Association for Computing Machinery, New York, NY, USA, 2022. URL: https://doi.org/10.1145/3526113.3545621. doi:10.1145/3526113.3545621.

[7] A. Yang, B. Yang, B. Hui, B. Zheng, B. Yu, C. Zhou, C. Li, C. Li, D. Liu, F. Huang, G. Dong, H. Wei, H. Lin, J. Tang, J. Wang, J. Yang, J. Tu, J. Zhang, J. Ma, J. Xu, J. Zhou, J. Bai, J. He, J. Lin, K. Dang, K. Lu, K. Chen, K. Yang, M. Li, M. Xue, N. Ni, P. Zhang, P. Wang, R. Peng, R. Men, R. Gao, R. Lin, S. Wang, S. Bai, S. Tan, T. Zhu, T. Li, T. Liu, W. Ge, X. Deng, X. Zhou, X. Ren, X. Zhang, X. Wei, X. Ren, Y. Fan, Y. Yao, Y. Zhang, Y. Wan, Y. Chu, Y. Liu, Z. Cui, Z. Zhang, Z. Fan, Qwen2 technical report, arXiv preprint arXiv:2407.10671 (2024).

[8] Q. Team, Qwen2.5: A party of foundation models, 2024. URL: https://qwenlm.github.io/blog/qwen2.5/.

[9] L. Heitz, R. Li, O. Inel, A. Bernstein, Informfully recommenders - reproducibility framework for diversity-aware intra-session recommendations, in: Proceedings of the Nineteenth ACM Conference on Recommender Systems, RecSys '25, Association for Computing Machinery, New York, NY, USA, 2025, p. 792–801. URL: https://doi.org/10.1145/3705328.3748148. doi:10.1145/3705328.3748148.