# MediaEval 2026 Grand Challenge: NewsImages

Sam Ziaie Kondori[1,*], Wong Chi Chong[1]

[1]National Tsing Hua University, Hsinchu, Taiwan

Abstract

This working note documents the approach of team TEAM_A++ to the MediaEval 2026 NewsImages task. We present a generation-and-selection pipeline that uses neither retrieval nor generator fine-tuning: for each headline, SDXL-Turbo produces $\boldsymbol{N}$ candidate thumbnails in a single diffusion step, an OCR filter removes those with spurious on-image text, and CFT-CLIP — a CLIP model fine-tuned on counterfactual news captions — selects the candidate best aligned with the article. Unlike prior work that applies CFT-CLIP only as a post-hoc score, we use it as the selection criterion within the generation loop. Offline, selection quality increases with $\boldsymbol{N}$ and CFT-CLIP discriminates more sharply than standard CLIP; in the crowdsourced evaluation, our run attained 3.041 against an editorial baseline of 2.880. These results indicate that, given an inexpensive one-step generator, thumbnail quality is more effectively obtained through candidate count and a story-specific selector than through heavier inference or fine-tuning.

**Keywords**

news thumbnails, image generation, diffusion models, CLIP, best-of-$N$ selection

## 1. Introduction

News thumbnails are a primary driver of reader engagement on online news platforms. Editors traditionally select a representative image for each article, but this breaks down for breaking news, where the original photographs may not yet exist and generic stock images often bear little relation to the story. Recent work shows that automatically retrieved or generated visuals can compete with, and sometimes outperform, editorial selection [1, 2]. Automated visuals, however, raise an ethical concern: a thumbnail must reflect the article accurately, without misleading readers into believing that it depicts a real event.

The NewsImages task at MediaEval 2026 [3] asks participants to recommend a fitting image for a news article from its text, using only open-source or open-weight components so the workflow is reproducible locally; we refer the reader to the overview paper for the full task, dataset, and evaluation protocol. Because the final evaluation is a crowdsourced human judgement of image fit, we treat that judgement as ground truth and any offline metric as a proxy, and frame the problem as generation followed by automated selection. We neither retrieve—exhaustive search over the $10^8$-image YFCC100M set [4] is infeasible and unlabeled subset selection is hard—nor fine-tune the generator, keeping the pipeline reproducible from public weights alone.

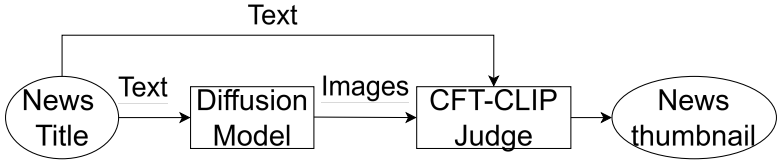


**Figure 1:** Overview of our news thumbnail generation system.

*MediaEval'26: Multimedia Evaluation Workshop, June 15–16, 2026, Amsterdam, the Netherlands and Online*
*Corresponding author.
samkondori@gapp.nthu.edu.tw (S. Z. Kondori); s112062121wong@gapp.nthu.edu.tw (W. C. Chong)

Fig. 1 summarizes our system: a distilled one-step diffusion model generates candidate thumbnails from a news title, and a CFT-CLIP judge [5] selects the candidate that best represents the article.

## 2. Related Work

Unlike the literal image–caption pairing assumed in image captioning [6], news-image *fit* is a matter of semantic alignment rather than direct correspondence. The NewsImages series studies this directly. The 2025 iteration [2] was the first to systematically compare retrieval, generation, and editorial selection, building on an earlier study of perceived text–image similarity [1]. A central finding of the 2026 overview [3] is that offline text–image similarity is only a weak predictor of the crowdsourced fit rating; the overview recommends scoring a set of candidates with an automated CLIP-based judge, which is the strategy our pipeline adopts.

The most closely related prior work is CFT-CLIP [5], a CLIP [7] model fine-tuned with counterfactual news text to assess whether a thumbnail is representative of an article. We adopt CFT-CLIP but use it differently: rather than applying it as a stand-alone measure after an image has been chosen, we make it the selection criterion within a best-of-$N$ generation loop, so that it determines which candidate is produced rather than only scoring the result. Standard CLIP serves as a baseline selector for comparison in Section 4.

## 3. Method

**Design Rationale** Two choices drive the design. First, the generator: a distilled one-step model such as SDXL-Turbo [8] produces an image in roughly $0.5$ s, about two orders of magnitude faster than a typical large diffusion model. This low per-sample cost makes best-of-$N$ sampling practical—we draw $N$ independent candidates and keep the best, so final quality is set by selection rather than the generator alone. Second, the selector: a thumbnail should be specific to its story, not merely a plausible illustration of the topic. Standard CLIP [7] rewards generic but visually similar images and is weak at this distinction; CFT-CLIP [5] is trained for it—it generates a counterfactual caption (e.g. by swapping the named entity) and scores an image lower against the counterfactual than against the true headline—so we use it as the selector and keep standard CLIP only as a baseline.

**Pipeline** The pipeline has four stages: (i) headline sanitization, (ii) candidate image generation, (iii) filtering, and (iv) selection. We first normalize the headline since dataset titles carry publisher names, section labels, and date fragments that provide little to no semantic value; we remove these while keeping any leading qualifiers such as "Fact check:" or "Report:" that change the meaning of the story. The cleaned title is then wrapped in a prompt that asks for an illustration rather than a photograph, in keeping with the task's preference for non-photorealistic images [3].

From this prompt, we draw $N$ candidate images using one-step diffusion generation. Because the generator occasionally renders spurious text onto an image, which the task penalizes, we apply an optical character recognition (OCR) pass and discard any candidates that contain legible text.

We then score each remaining candidate against the article text with CFT-CLIP and return the highest-scoring one,

$$\hat{x} = \arg\max_{x \in \mathcal{C}} \; s_{\text{CFT}}(x, t), \tag{1}$$

where $\mathcal{C}$ is the set of candidates that survive the filter for headline $t$ and $s_{\text{CFT}}$ is the CFT-CLIP image–text similarity. The selected image $\hat{x}$ is the submitted thumbnail; we also record each candidate's standard CLIP score for the comparison in Section 4. Model versions, the exact prompt template, and all thresholds are documented with our released code.

## 4. Experiments and Results

**Experimental Setup** All development experiments use real headlines sampled from the NewsImages training set of 8,500 English-language articles collected from GDELT during 2022–2023 [3]; the final submitted run covers the 800 articles of the test set. We generate with SDXL-Turbo [8] and select with CFT-CLIP [5], while also using standard CLIP [7] as the baseline selector. The official metric is the crowdsourced 5-point image-fit rating; offline, we report the CFT-CLIP score to compare the two selectors (Exp. 1) and the effect of candidate count $N$ (Exp. 2).

**Exp. 1: CLIP vs. CFT-CLIP as Selector** We compare the two metrics over a fixed pool of $N = 20$ candidates per headline across 30 headlines.

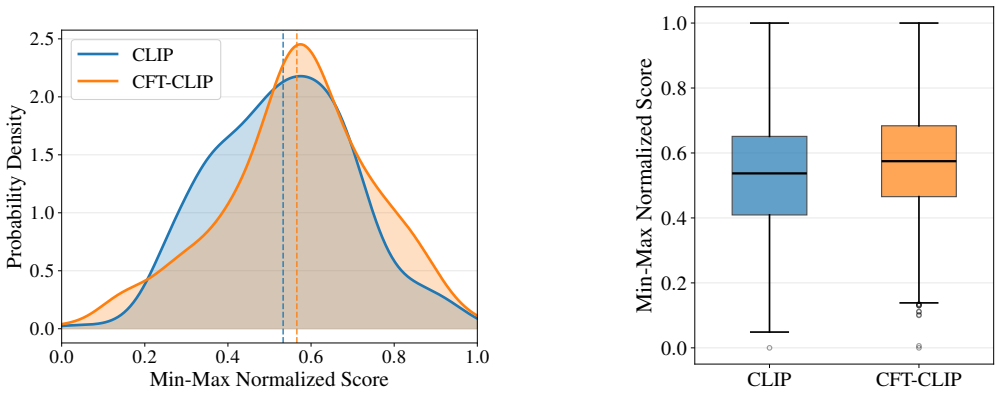


**Figure 2:** (a) Score distributions for standard CLIP and CFT-CLIP across all candidates. Each metric is min-max normalized independently to $[0, 1]$; dashed lines indicate their respective means. (b) Boxplot of the same normalized scores.

Fig. 2 shows that CFT-CLIP scores center at a higher normalized level than standard CLIP, reflecting the two models' different training objectives. Standard CLIP is a general-purpose vision–language model that rewards visual similarity to the query; CFT-CLIP is fine-tuned to penalize images that fit a counterfactual version of the headline nearly as well as the true one, making it sensitive to story-specific content. The practical consequence is shown in Fig. 3: for the same candidate pool, CFT-CLIP selects an image that depicts the specific event described in the headline, while CLIP selects an image that is topically related but not story-specific.

**Exp. 2: Best-of-N Scaling** We measure the mean CFT-CLIP score of the selected thumbnail as the pool size grows, $N \in \{1, 5, 25, 125\}$.

Fig. 4 shows that the mean score increases monotonically with $N$ but with diminishing returns. Drawing more candidates and keeping the best therefore improves alignment without any change to the model, at the cost of additional generation. Because that cost grows linearly with $N$ while the gain shrinks, we submit at $N = 5$ as a compromise between fit and generation cost.

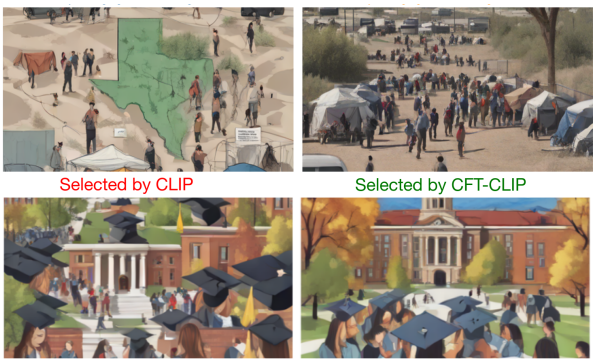


**Figure 3:** Candidates chosen by CLIP vs. CFT-CLIP for the same headline. Top: *DOJ Assessing Migrant Treatment along Texas Border.* Bottom: *Career Advance Colorado brings free tuition to select programs at 19 state colleges this fall.*

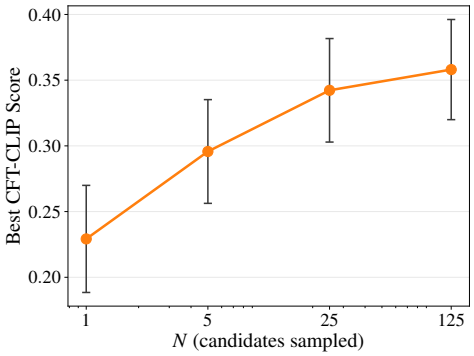


**Figure 4:** Mean CFT-CLIP score of the selected thumbnail vs. candidate count $N$ (log scale; error bars: 95% CI, $n = 30$ articles).

**Submitted Run** For the official submission (group TEAM_A++) we generate one thumbnail for each of the 800 test articles with the full pipeline at $N = 5$.

**Table 1**
Crowdsourced image-fit ratings (higher is better).

| Run / Baseline | Mean image-fit rating |
|---|---|
| Editorial original image (baseline) | 2.880 |
| NewsImages 2025 top run: CVG-IBA [9] | 3.400 |
| Ours, TEAM_A++ | 3.041 |

Table 1 reports the mean image-fit ratings. Our run improves on the editorial-original baseline (3.041 vs. 2.880) but does not match CVG-IBA, the top-rated NewsImages 2025 submission (3.400).

## 5. Analysis and Conclusion

Best-of-$N$ selection raised the offline alignment score monotonically with $N$, and the submitted run improved on the editorial baseline (3.041 vs. 2.880) while trailing the best run from 2025 [9], indicating that a cheap one-step generator paired with a story-specific selector is a viable route to fit, but not yet a winning one.

Our approach has three main limitations. SDXL-Turbo has no world knowledge of specific people or places, so headlines that name individuals tend to produce generic or misidentified likenesses. The one-step generator also trades sample diversity for speed and occasionally produces degenerate candidates even at moderate $N$. Finally, because CFT-CLIP is both our selector and our offline proxy, a high automated score does not guarantee a high crowdsourced fit rating; the human evaluation in Section 4 is the real test of the method.

Several directions could address these limitations. A stronger open-weight text-to-image model, such as the distilled versions of Z-Image or FLUX, would improve named-entity grounding and sample quality. Category-specific prompt templates could better match the conventions of different news topics. A hybrid step could also retrieve a real image when a headline names a well-known entity, falling back to generation otherwise.

## Declaration on Generative AI

During the preparation of this work, the authors used Grammarly for grammar and spelling checking and proofreading. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the manuscript's content.

## References


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

[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, 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.

[4] B. Thomee, D. A. Shamma, G. Friedland, B. Elizalde, K. Ni, D. Poland, D. Borth, L.-J. Li, The new data and new challenges in multimedia research, arXiv preprint arXiv:1503.01817 1 (2015).

[5] Y. Yoon, S. Yoon, K. Park, Assessing news thumbnail representativeness: Counterfactual text can enhance the cross-modal matching ability, in: Findings of the Association for Computational Linguistics: ACL 2024, 2024, pp. 9009–9024.

[6] N. Oostdijk, H. van Halteren, E. Başar, M. Larson, The connection between the text and images of news articles: New insights for multimedia analysis, in: Proceedings of the 12th Language Resources and Evaluation Conference (LREC), 2020, pp. 4343–4351.

[7] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PmLR, 2021, pp. 8748–8763.

[8] A. Sauer, D. Lorenz, A. Blattmann, R. Rombach, Adversarial diffusion distillation, in: European Conference on Computer Vision, Springer, 2024, pp. 87–103.

[9] M. Khan, A. Subhani, M. Rafi, A. Tahir, Balancing relevance and compliance: Text-to-image methods for news articles visualization, in: Working Notes Proceedings of the MediaEval 2025 Workshop, 2025.