Finding Representative Images for Wikipedia by Quizzing VLMs

Tenghao JI
University of Michigan

Eytan Adar
University of Michigan

Abstract

Images play a vital role in improving the readability and comprehension of text (e.g., Wikipedia articles) by serving as "illustrative aids." However, not all images are equally effective and not all editors are trained in their selection. We propose QuizRank, a method that leverages LLMs and VLMs to rank candidate images by how well they support answering multiple-choice questions about key visual characteristics. To better discriminate visually similar items, we introduce *Contrastive QuizRank*, which generates questions from differences between target and distractor topics (e.g., *Western Bluebird* vs. *Mountain Bluebird*). We show high agreement with human quiz-takers and improved ranking quality for Wikipedia-style image selection.

Keywords: Image search, Representative images, Image ranking

Introduction

In contexts such as Wikipedia—where images and text combine—a key function of images is to be illustrative and representative. Editors are guided to select "images [that] look *like* what they are meant to illustrate" (emphasis in original) and to serve as an "illustrative aid" (Wikimedia, 2025). Images can improve comprehension and engagement (Clark and Lyons, 2010; Guo et al., 2020; Carney and Levin, 2002), yet finding good images is difficult: as of January 2026 there are over 7 million English Wikipedia articles and 132.5 million images on Wikimedia Commons. Selecting bad images may result in illustrative aids that are generic, unclear, or mismatched to the article (Rama et al., 2022; Navarrete and Villaespesa, 2020).

Prior work mainly relies on caption-based text–image similarity (Agrawal et al., 2011; Leake et al., 2020), representative sampling from an image set (Han et al., 2020), or prompting LLMs/VLMs to judge fit (Li et al., 2024; Yang and Lin, 2024). These approaches can be coarse and may not reflect the intent of an image as an illustrative aid. To address this, we propose QuizRank (Figure 1). QuizRank treats image selection as a learning assessment: an image is better if it helps a model answer questions about the topic.

QuizRank (Figure 1) analyzes article text to identify visual characteristics and converts them into multiple-choice questions (1a). A VLM then answers these questions conditioned on each candidate image (2); images that better support answering are ranked higher (3). To better discriminate visually similar candidates, QuizRank can generate quizzes contrastively (1b) by leveraging reasonable distractors (e.g., *Gujia* vs. *Chandrakala*), drawing inspiration from prior work on visual disambiguation (Leggett and Kirchoff, 2011).

We evaluate QuizRank by comparing VLM and human performance on answering the same questions given different images, showing strong correspondence in image scoring and in/out-of-class discrimination. We further compare QuizRank against alternatives (e.g., CLIP-based similarity) using both closed and open-source models, and demonstrate effective image selection across diverse Wikipedia topics.

QuizRank

We define high-quality illustrative aids as images that are strong *in-class* representatives (they depict key visual features described in the article) and strong *between-class* representatives (they help distinguish the target from related topics). To find these, QuizRank proceeds in three main steps (Figure 1). (1a) It generates multiple-choice questions (MCQs) from the topic's Wikipedia page, prioritizing visually grounded attributes (e.g., material, texture, structure) and avoiding topic-identity leakage using generalized phrasing (inspired by MCQ design guidance (Haladyna, 2004)). For example, if an article identifies that a key feature of a statue is that it is bronze, the question might be "what material is the statue made from? A) Bronze, B) Marble, C) Copper." Each image (including possible distractors) are fed to a VLM with the MCQ 'test' (2). Again, the intuition is that if the VLM can better answer a question given the image as an 'aid,' that image is better. The images can then be ranked according to how many questions they helped the VLM answer (3). For some cases, a distractor becomes 'blended' in the ranking. This indicates that a two topics have similar visual characteristics. To address this, we provide a

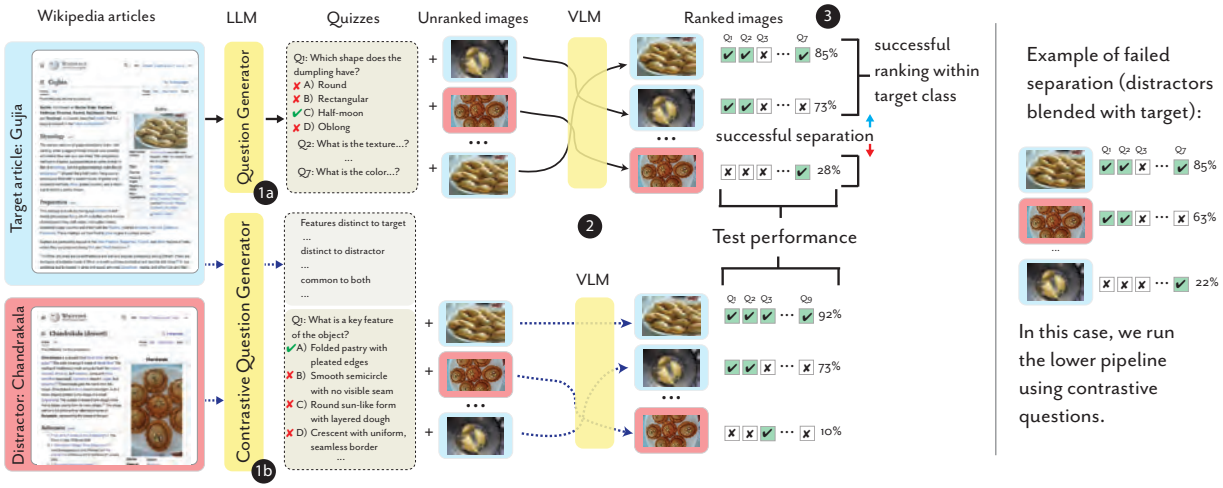

Figure 1: Example of QuizRank with standard (top) and contrastive ranking (bottom). A quiz is generated from the target description (and, for contrastive, one or more distractors). The VLM answers questions per image and the resulting scores induce a ranking.

Constrastive version of QuizRank (1b) that first identifies which features are unique to the target concept relative to the distractor. Questions are only generated on the basis of these unique elements. We score and rank images by quiz performance (e.g., normalized fraction correct), retaining the image–question correctness matrix to surface complementary bundles when no single image covers all properties (model-as-judge setup (Li et al., 2024)).

Results

We evaluate QuizRank along two goals: (i) whether VLM quiz performance aligns with human quiz performance when observing the same image, and (ii) whether QuizRank ranks images better than a CLIP-similarity baseline. Additional information (prompt details, data, analyses, and an example interface call QuizRankSearch) are available on our supplement website[1].

Data and procedure. We evaluate 40 Wikipedia (a subset of those in (Silva et al., 2024)) with four candidate images per topic (three in-class and one distractor). For each topic, we use QuizRank to generate MCQs and score each image. We also collect human quiz scores on the same questions and images via a Prolific crowd study (320 participants generating 10 scores per image).

Human–VLM correspondence. Figure 2 compares per-image accuracy between humans and the VLM. We observe strong alignment for both target and distractor images (Pearson $r = 0.6125$, $p < .001$), suggesting the questions capture image-evident visual properties rather than purely textual cues. We also find that the VLM's absolute accuracy is often higher than humans', which may reflect differences in attention or inferential ability; however, because all candidate images are graded by the same model, ranking remains meaningful for selection. Statistical tests (Table 1) further support these findings. Both humans and the VLM perform significantly better on non-distractor images compared to distractors. The VLM consistently outperforms humans on non-distractor images, while differences on distractor images are marginal and not statistically significant.

CLIP similarity baseline. We rank images using CLIP similarity between each image and a short visual description of the topic extracted from the Wikipedia page. When comparing ranks to human quiz-based ranks, CLIP achieves a weak mean Spearman correlation of 0.305, while QuizRank achieves 0.535 on the same topics. This suggests that caption-style similarity can be too coarse for selecting representative images, whereas quiz-based evaluation better isolates whether an image provides evidence for specific visual characteristics.

Discussion/Conclusions

QuizRank offers a new method for selecting relevant images for Wikipedia articles, evaluating images based on the ability to answer visual questions. The contrastive question generation helps distinguish similar images. Future work may allow us to expand the approach to additional topics (e.g., abstract or biographical) and to other language domains. QuizRank can be used in search interfaces to aid editors in finding good images for articles or sections (see the supplement for a possible interface). Our evaluations demonstrate that the approach is scalable and robust–identifying good, representative, illustrative aids.

[1] https://anonymous.4open.science/r/quizranksearch-supplement-117C/

References

[Agrawal et al.2011] Rakesh Agrawal, Sreenivas Gollapudi, Anitha Kannan, and Krishnaram Kenthapadi. 2011. Enriching textbooks with images. In *CIKM'11*.

[Carney and Levin2002] Russell N Carney and Joel R Levin. 2002. Pictorial illustrations still improve students' learning from text. *Educational psychology review*, 14(1):5–26.

[Clark and Lyons2010] Ruth C Clark and Chopeta Lyons. 2010. *Graphics for learning: Proven guidelines for planning, designing, and evaluating visuals in training materials*. John Wiley & Sons.

[Guo et al.2020] Daibao Guo, Shuai Zhang, Katherine Landau Wright, and Erin M McTigue. 2020. Do you get the picture? a meta-analysis of the effect of graphics on reading comprehension. *AERA*, 6(1).

[Haladyna2004] Thomas M Haladyna. 2004. *Developing and validating multiple-choice test items*. Routledge.

[Han et al.2020] Shanshan Han, Fu Ren, Qingyun Du, and Dawei Gui. 2020. Extracting representative images of tourist attractions from flickr by combining an improved cluster method and multiple deep learning models. *Int. J. of Geo-Info.*, 9(2):81.

[Leake et al.2020] Mackenzie Leake, Hijung Valentina Shin, Joy O Kim, and Maneesh Agrawala. 2020. Generating audio-visual slideshows from text articles using word concreteness. In *CHI'20*, pages 1–11.

[Leggett and Kirchoff2011] Roxanne Leggett and Bruce K. Kirchoff. 2011. Image use in field guides and identification keys: review and recommendations. *AoB PLANTS*, 2011, 01.

[Li et al.2024] Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, et al. 2024. From generation to judgment: Opportunities and challenges of llm-as-a-judge. *arXiv:2411.16594*.

[Navarrete and Villaespesa2020] Trilce Navarrete and Elena Villaespesa. 2020. Image-based information: paintings in wikipedia. *J. of Documentation*.

[Rama et al.2022] Daniele Rama, Tiziano Piccardi, Miriam Redi, and Rossano Schifanella. 2022. A large scale study of reader interactions with images on wikipedia. *EPJ Data Science*, 11(1):1.

[Silva et al.2024] Anita Silva, Maria Tracy, Katharina Reinecke, Eytan Adar, and Miriam Redi. 2024. Imagine a dragon made of seaweed: How images enhance learning in wikipedia. *arXiv:2403.07613*.

[Wikimedia2025] Wikimedia. 2025. Wikipedia:manual of style/images. `https://w.wiki/6HT7`.

[Yang and Lin2024] Jheng-Hong Yang and Jimmy Lin. 2024. Toward automatic relevance judgment using vision–language models for image–text retrieval evaluation. *arXiv:2408.01363*.

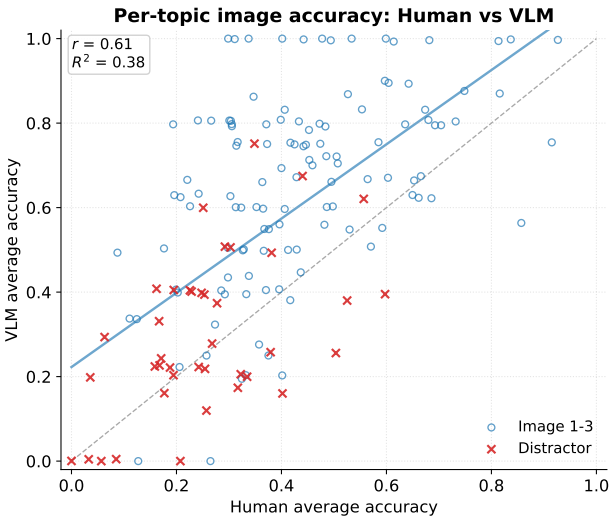

Figure 2: Per-image average accuracy comparison between human participants and the VLM. Red points represent distractor images. The dotted line indicates parity.

Comparison	n1	n2	KW	p	ANOVA	p
Human non vs distr	120	40	30.420	0.0000	33.717	0.0000
VLM non vs distr	120	40	53.331	0.0000	83.027	0.0000
Human vs VLM (non)	120	120	57.726	0.0000	70.836	0.0000
Human vs VLM (distr)	40	40	0.753	0.3857	1.155	0.2858

Table 1: Statistical test results comparing human and VLM performance on different image types.