# Toward Automatic Relevance Judgment using Vision–Language Models for Image–Text Retrieval Evaluation

Jheng-Hong Yang[1], Jimmy Lin[1]

[1]University of Waterloo, Canada

**Abstract**

Vision–Language Models (VLMs) have demonstrated success across diverse applications, yet their potential to assist in relevance judgments remains uncertain. This paper assesses the relevance estimation capabilities of VLMs, including CLIP, LLaVA, and GPT-4V, within a large-scale *ad hoc* retrieval task tailored for multimedia content creation in a zero-shot fashion. Preliminary experiments reveal the following: (1) Both LLaVA and GPT-4V, encompassing open-source and closed-source visual-instruction-tuned Large Language Models (LLMs), achieve notable Kendall's $\tau \sim 0.4$ when compared to human relevance judgments, surpassing the CLIPScore metric. (2) While CLIPScore is strongly preferred, LLMs are less biased towards CLIP-based retrieval systems. (3) GPT-4V's score distribution aligns more closely with human judgments than other models, achieving a Cohen's $\kappa$ value of around 0.08, which outperforms CLIPScore at approximately -0.096. These findings underscore the potential of LLM-powered VLMs in enhancing relevance judgments.

**Keywords**

Relevance Assessments, Image–Text Retrieval, Vision–Language Model, Large Language Model

## 1. Introduction

Cranfield-style test collections, consisting of a document corpus, a set of queries, and manually assessed relevance judgments, have long served as the foundation of information retrieval research [1]. However, evaluating every document for every query in a substantial corpus often proves cost-prohibitive. To tackle this challenge, a subset of documents is selected for assessment through a pooling process. While this method is cost-effective compared to user studies, it has limitations due to its simplifications and struggles to adapt to complex search scenarios and large document collections.

In this study, we explore the adaptability of model-based relevance judgments for image–text retrieval evaluation. Leveraging model-based retrieval judgments presents an appealing option. Not only does it provide valuable insights before undertaking the laborious processes of document curation, query creation, and costly annotation, but it also has the potential to extend and scale up to complex search scenarios and large document collections. To explore opportunities and meet the demands for large-scale, fine-grained, and long-form text enrichment scenarios in image-text retrieval evaluation [2, 3, 4, 5], our objective is to extend the human-

*LLM4Eval: The First Workshop on Large Language Models for Evaluation in Information Retrieval, 18 July 2024, Washington DC, United States*

jheng-hong.yang@uwaterloo.ca (J. Yang); jimmylin@uwaterloo.ca (J. Lin)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

CEUR-WS.org/Vol-3752/paper7.pdf

CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073

machine collaborative framework proposed by Faggioli et al. [6] to the context of image-text retrieval evaluation, alongside widely adopted model-based image-text evaluation metrics [7, 8, 9, 10, 11].

Our primary focus is on a fully automatic evaluation paradigm, where we harness the capabilities of Vision–Language Models (VLMs), including CLIP [12], as well as visual instruction-tuned Large Language Models (LLMs) like LLaVA [13, 14] and GPT-4V [15]. To evaluate this approach, we conducted a pilot study using the TREC-AToMiC 2023 test collection, which is designed for multimedia content creation [5], based on our instruction prompt template for VLMs (cf. Table 1 and Section 3.2).

We observe that model-based relevance judgments generated by visual instruction-tuned LLMs outperform the widely adopted CLIPScore [7] in terms of ranking correlations and agreements when compared to human annotations. While this discovery holds promise, we also uncover the potential evaluation bias when using model-based relevance judgments. Our analysis reveals a bias in favor of CLIP-based retrieval systems in the rankings when employing model-based relevance judgments, resulting in higher overall effectiveness assessments for these systems. In summary, our contributions can be distilled as follows:

- We demonstrate and explore the feasibility of incorporating VLMs for fully automatic image–text retrieval evaluation.
- We shed light on the evaluation bias when utilizing model-based relevance judgments.

## 2. Related Work

**Evaluation Metrics for Image–Text Relevance.** Nowadays, model-based evaluation metrics are widely utilized in various vision–language tasks, including image captioning [7, 16] and text-to-image synthesis [8, 17]. Among model-based approaches, CLIP-based methods [8, 9, 18, 10, 11], such as CLIPScore [7], are particularly prevalent. However, while these metrics are capable of measuring coarse text-image similarity, they may fall short in capturing fine-grained image–text correspondence [3, 19]. Recent research has highlighted the effectiveness of enhancing model-based evaluation metrics by leveraging LLMs to harness their reasoning capabilities [16, 20, 21]. There exists significant potential for incorporating LLMs into model-based approaches, as LLM outputs are not limited to mere scores but can also provide free-form texts, e.g., reasons, for further analysis and many downstream tasks [22].

**Model-based Relevance Judgments.** Traditionally, relevance judgments in retrieval tasks have adhered to the Cranfield evaluation paradigm due to its cost-effectiveness, reproducibility, and reliability when compared to conducting user studies. However, this approach often relies on simplified assumptions and encounters scalability challenges. Researchers have recently explored model-based automatic relevance estimation as a promising alternative. This approach aims to optimize human-machine collaboration to obtain ideal relevance judgments. Notably, studies of Dietz and Dalton [23] and Faggioli et al. [6] have revealed high rank correlations between model-based and human-based judgments. Additionally, MacAvaney and Soldaini [24] have delved into the task of filling gaps in relevance judgments using model-based annotations.

**Table 1**
Prompt template for relevance estimation. The VLMs are expected to take text $q$ and image $d$ independently. The prompts are only applied to the textual input $q$, while the VLMs process the pixel values of image $d$ directly.

**Text Input:**

**Context:**
Page Title: <page_title>
Page Context: <page_context>
Section Title: <section_title>
Section Context: <section_context>
**Relevance Instruction:**
Think carefully about which images best illustrate the SECTION subject matter. Given the text and the image please answer the following questions given the criteria listed as follows:
* Images must be significant and relevant in the topic's context, not primarily decorative. They are often an important illustrative aid to understanding.
* Images should look like what they are meant to illustrate, whether or not they are provably authentic.
* Textual information should almost always be entered as text rather than as an image.
**Output Instruction:**
Relevance: Rate the image's overall relevance (integer, scale: 1-100) in terms of matching the text.
Output format should be: "Relevance: <score>"

**Image Input:**

## 3. Methodology

In this study, we investigate techniques for estimating image-text relevance scores, denoted as $\mathcal{F}(q, d) \in \mathbb{R}$, where $q$ represents the text (query) and $d$ represents the image (document). Our primary focus is on utilizing VLMs to generate relevance scores, akin to empirical values annotated by human assessors denoted as $\hat{\mathcal{F}}(q, d)$. The main objective is to assess the proximity between model-based $\mathcal{F}$ and human-based $\hat{\mathcal{F}}$ in image–text retrieval evaluation. We begin with a discussion of the setting for human-based annotations, followed by the process for generating model-based annotations.

### 3.1. Human-based Annotations

Our primary focus revolves around a critical aspect of multimedia content creation, specifically, the image suggestion task, an *ad hoc* image retrieval task as part of the AToMiC track in the TREC conference 2023 (TREC-AToMiC 2023).[1] The image suggestion task aims to identify relevant images from a predefined collection, given a specific section of an article. Its overarching goal is to enrich textual content by selecting images that aid readers in better comprehending the material.

Relevance scores for this task are meticulously annotated by NIST assessors, adhering to the

[1]https://trec-atomic.github.io/trec-2023-guidelines

TREC-style top-$k$ pooling relevance annotation process. A total of sixteen valid participant runs, generated by diverse image–text retrieval systems, are considered, encompassing (CLIP-based) dense retrievers, learned sparse retrievers, caption-based retrievers, hybrid systems, and multi-stage retrieval systems. The pooling depth is set to 25 for eight baseline systems and 30 for the remaining participant runs.

NIST assessors classify candidate results into three graded relevance levels to capture nuances in suitability, guided by the content of the test query. The test query comprises textual elements such as the section title, section context description, page title, and page context description. Assessors base their relevance judgments on the following criteria:

- **0 (Non-relevant)**: Candidates deemed irrelevant.
- **1 (Related)**: Candidates that are related but not relevant to the section context are categorized as related. They contain pertinent information but do not align with the section's context.
- **2 (Relevant)**: These candidates are considered relevant to the section context and effectively illustrate it.

### 3.2. Model-based Annotations

For automatic relevance estimation, we employ pretrained VLMs as our relevance estimator, denoted as $\mathcal{F}(q, d \mid \mathcal{P})$. Our relevance estimator produces relevance scores given a pair of $q$ and $d$, which is conditioned on $\mathcal{P}$, where $\mathcal{P}$ represents the prompt template we used to instruct the models. Prompt engineering is a commonly adopted technique for enhancing or guiding VLMs and LLMs in various tasks [25, 12]. It's important to note that our current focus is on pointwise estimation, leaving more advanced ranking methods (such as pairwise or listwise) that consider multiple $q$ and $d$ for future exploration [26, 27].

**Prompt Template Design** In line with our approach to relevance score annotation, we have created a prompt template designed to guide models in generating relevance scores. The prompt template, presented in Table 1, has been constructed based on our heuristics and is not an exhaustive search of all possible templates. Pretrained VLMs are expected to take both $q$ and $d$ to produce a relevance score following the instructions defined in the prompt template $\mathcal{P}$. We anticipate that VLMs will independently process textual and visual information, and our prompt template is only applied to textual inputs.Our template comprises three essential components:

- Context: This section processes the textual information from $q$.[2]
- Relevance Instruction: It incorporates task-specific information designed to provide VLMs with an understanding of the task.
- Output Instruction: This component offers instructions concerning the expected output, e.g., output types and format.

[2]For VLMs with limited context windows, e.g., CLIP, we only take the texts in the context part and ignore all the rest instructions.

**From Scores to Relevance Judgments.** We utilize parsing scripts to process the relevance scores generated by the models and convert them into relevance judgments.[3] Considering potential score variations across different models, we apply an additional heuristic rule to map these scores into graded relevance levels: 0 (non-relevant), 1 (related), and 2 (relevant). Specifically, scores falling below the median value are categorized as 0; scores within the 50-75th quantile range are designated as 1; and scores exceeding the 75th quantile are assigned a relevance level of 2.

**Table 2**
Ranking correlation and judgment agreement analysis. Correlations are reported in terms of Kendall's $\tau$, Spearman's $\rho_s$, and Pearson's $\rho_p$, whereas judgment agreement is reported in terms of Cohen's $\kappa$ when comparing to NIST qrels.

| | | NDCG@10 | | | MAP | | | Agreement |
|---|---|---|---|---|---|---|---|---|
| Model | Version | $\tau$ | $\rho_s$ | $\rho_p$ | $\tau$ | $\rho_s$ | $\rho_p$ | $\kappa$ |
| **CLIP-S** | `openai/clip-vit-large-patch14` | 0.200 | 0.253 | 0.209 | 0.333 | 0.356 | 0.418 | -0.096 |
| **LLaVA-7b** | `v1.5-7b` | 0.400 | 0.532 | 0.633 | 0.483 | 0.597 | 0.507 | -0.003 |
| **LLaVA-13b** | `v1.5-13b` | 0.433 | 0.559 | 0.659 | 0.517 | 0.618 | 0.523 | 0.010 |
| **GPT-4V** | `1106-vision-preview` | 0.400 | 0.544 | 0.540 | 0.500 | 0.594 | 0.470 | 0.080 |

# 4. Experiments

We have undertaken an empirical comparison between human assessors and vision-language models to offer an initial evaluation of their current capabilities in estimating relevance judgments. This comparative analysis encompasses one embedding-based model (CLIP) and two LLMs trained by visual instruction tuning (LLaVA and GPT-4V). The experiments were carried out in January 2024.

## 4.1. Setups

**Test Collection.** Our study focuses on the image suggestion task in TREC-AToMiC 2023. In this task, queries are sections from Wikipedia pages, and the corpus contains images from Wikipedia. We assess VLMs' ability to assign relevance labels to 9,818 image–text pairs across 74 test topics. We predict relevance scores, generate qrels for 16 retrieval runs, and compare them with NIST human-assigned qrels. Note that the test topics consist of Wikipedia text sections (level-3 vital articles) without accompanying images, and NIST qrels are not publicly accessible during the training of VLMs we study in this work.

**Vision–Language Models.** Our experiments feature three models: CLIP [12], LLaVA [13, 14], and GPT-4V [15]. CLIP serves as a versatile baseline model, offering similarity scores for image–text pairs. We use CLIPScore [7] (referred to as CLIP-S) for calculating relevance with CLIP. However, CLIP has limitations due to its text encoder's token limit (77 tokens), making it less adaptable for complex tasks with lengthy contexts. In contrast, LLMs like LLaVA and GPT-4V,

[3]For CLIP, relevance scores are computed using text and image embeddings directly.

**Table 3**
Evaluation bias assessment using Relative $\Delta$ in terms of NDCG@10 and MAP. A positive $\Delta$ favors CLIP-based systems, while a negative $\Delta$ favors other types of systems.

| Model | $\Delta$(NDCG@10) | $\Delta$(MAP) |
|---|---|---|
| **CLIP-S** | **114.7** | **120.5** |
| **LLaVA-7b** | 58.5 | 86.6 |
| **LLaVA-13b** | 55.8 | 83.1 |
| **GPT-4V** | 64.0 | 91.3 |
| **Human** | -11.7 | -19.5 |

fine-tuned for visual instruction understanding, possess larger text encoders capable of handling extended context. These models excel in various vision-language tasks, making them more versatile compared to CLIP.

## 4.2. Correlation Study

In this subsection, our primary aim is to investigate the extrinsic properties of relevance judgments generated by various approaches, where we base our analysis on retrieval runs and ranking metrics. While various techniques exist to enhance the capabilities of vision-language models, including prompt engineering, few-shot instructions, and instruction tuning, our current focus centers on examining their zero-shot capabilities. We defer the exploration of other methods to future research endeavors. Following the work of Voorhees [28], we undertake an investigation into the system ranking correlation and the agreement between the relevance labels estimated by the model and those provided by NIST annotators. We evaluate the ranking correlations concerning the primary metrics utilized in the AToMiC track: NDCG@10 and MAP, and calculate Kendall's $\tau$, Spearman's $\rho_s$, and Pearson's $\rho_p$. In our agreement study, we compute Cohen's $\kappa$ using NIST's qrels as references.

**Overall.** The primary results are showcased in Table 2, where rows correspond to the backbone model used for relevance judgment generation. Notably, models leveraging LLMs such as LLaVA and GPT-4V outperform the CLIP-S baseline concerning ranking correlation. Specifically, they achieve Kendall's $\tau$ values of approximately 0.4 for NDCG@10 and around 0.5 for MAP. For comparison, previous research reported 0.9 for $\tau$ for MAP when comparing two types of human judgments [28]. While there is still room for further improvement, our observations already demonstrate enhancement compared to the CLIP-S baseline: 0.200 (0.333) for NDCG@10 (MAP). Moreover, other correlation coefficients, including Spearman and Pearson, corroborate the trends identified by Kendall's $\tau$. Additionally, we notice a rising trend in agreement levels when transitioning from CLIP-S (-0.096) to GPT-4V (0.080), as evidenced by Cohen's $\kappa$ values. The agreements achieved by the two largest models (LLaVA-13b and GPT-4V) are categorized as 'slight,' which represents an improvement over the smaller LLaVA-7b model and the baseline.

**Evaluation Bias** Model-based evaluations can introduce bias, often favoring models that are closely related to the assessor model [29, 30]. We term this phenomenon as *evaluation bias*.

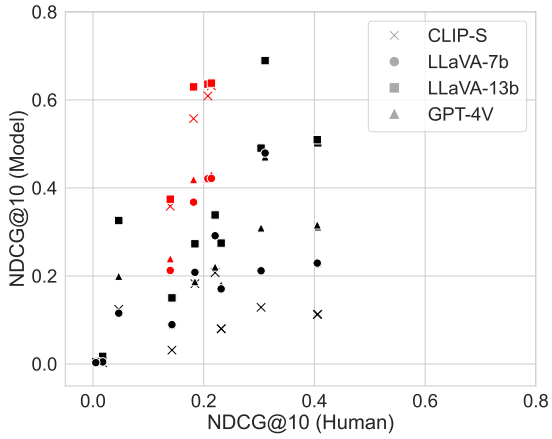


**Figure 1:** Scatter plots of effectiveness (NDCG@10) for TREC-AToMiC 2023 runs using human-based and model-based qrels. Each data point represents the mean effectiveness of a single run evaluated with different qrels. CLIP-based runs are highlighted in red. Best viewed in color.

This is distinct from *source bias* which indicates that neural retrievers might prefer contents generated by generative models [31]. To address this potential concern, we conducted an initial analysis using the scatter plot presented in Fig. 1. In this analysis, we compared the NDCG@10 scores of the 16 submissions made by participants employing different sets of qrels. Each data point on the plot corresponds to a specific run, with distinct markers representing variations in results based on relevance estimation models. Upon closer examination of the plot, we identified a positive correlation between model-based and human-based qrels. Notably, the effectiveness of submitted systems appeared slightly higher when compared to those using human-based qrels.

To gain deeper insights, we've visually highlighted CLIP-based submissions in red for a thorough investigation. This visual distinction underscores the preference for model-based qrels for CLIP-based systems, especially evident with CLIP-S qrels. We quantitatively assess this bias using a metric adapted from the work of Dai et al. [31]:

$$\text{Relative } \Delta = 2\frac{\text{Metric}_{\text{CLIP-based}} - \text{Metric}_{\text{Others}}}{\text{Metric}_{\text{CLIP-based}} + \text{Metric}_{\text{Others}}} \times 100\%,$$

here Metric stands for a measure, e.g., NDCG@$k$, averaged across systems. Observing Table 3, CLIP-S exhibits a strong bias, with Relative $\Delta = 114.7$ for NDCG@10 and 120.5 for MAP. LLM-based approaches also display a slight bias towards CLIP-based systems, possibly because both LLaVA and GPT-4V rely on CLIP embeddings for image representations. In contrast, human-based qrels show the lowest bias: -11.7 for NDCG@10 and -19.5 for MAP.

### 4.3. Estimated Relevance Analysis

In this subsection, we aim to explore the intrinsic properties of relevance judgments generated by various systems. We began our analysis by examining score distributions, visualized in

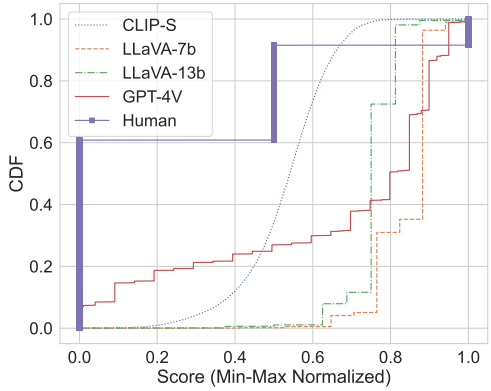


**Figure 2:** Cumulative distribution function (CDF) plot of relevance scores from various models. Human stands for relevance annotations of NIST qrels.

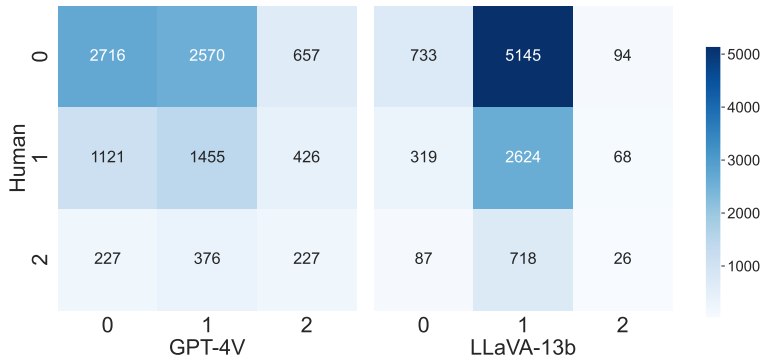


**Figure 3:** Confusion matrices comparing human-based and model-based qrels. Tick labels 0/1/2 represent Non-relevant/Related/Relevant graded levels. Best viewed in color.

Figures 2 and 3, to gain insights into model-based scores.

Figure 2 presents a Cumulative Distribution Function (CDF) plot of scores before post-processing into relevance levels (0, 1, and 2). We included NIST qrels (human) results for reference. Notably, GPT-4V's score distribution closely aligns with the human CDF, while CLIP-S exhibits a smoother S-shaped distribution with limited representation of low-relevance data. LLaVA produces tightly concentrated scores, adding complexity to post-processing, particularly when compared to GPT-4V.

Figure 3 illustrates confusion matrices, highlighting LLaVA's tendency to generate more 1 (related) judgments, fewer 2 (relevant), and 0 (non-relevant) judgments compared to GPT-4V. We anticipate that future models will strive to produce score distributions that better match human annotations, thereby addressing these challenges and limitations. Further studies [32] on harnessing LLMs' relevance prediction capability are necessary.

## 5. Conclusion

This study delves into the capabilities of VLMs such as CLIP, LLaVA, and GPT-4V for automating relevance judgments in image–text retrieval evaluation. Our findings reveal that visual-instruction-tuned LLMs outperform traditional metrics like CLIPScore in aligning with human judgments, with GPT-4V showing particular promise due to its closer alignment with human judgment distributions.

Despite these advancements and low cost of model-based relevance annotation,[4] challenges such as evaluation bias and the complexity of mimicking human judgments remain. These issues underscore the need for ongoing model refinement and exploration of new techniques to enhance the reliability and scalability of automated relevance judgments.

In conclusion, our research highlights the potential of VLMs in streamlining multimedia content creation while also pointing to the critical areas requiring further investigation. The path toward fully automated relevance judgment is complex, necessitating continued collaborative efforts in the research community to harness the full potential of VLMs in this domain.

## Acknowledgements

This research was supported in part by the Canada First Research Excellence Fund and the Natural Sciences and Engineering Research Council (NSERC) of Canada.

## References

[1] C. W. Cleverdon, The aslib cranfield research project on the comparative efficiency of indexing systems, in: Aslib Proceedings, volume 12, 1960, pp. 421–431.

[2] F. Schneider, Ö. Alaçam, X. Wang, C. Biemann, Towards multi-modal text-image retrieval to improve human reading, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop, 2021.

[3] E. Kreiss, C. Bennett, S. Hooshmand, E. Zelikman, M. Ringel Morris, C. Potts, Context matters for image descriptions for accessibility: Challenges for referenceless evaluation metrics, in: Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, 2022, pp. 4685–4697.

[4] J. Singh, V. Zouhar, M. Sachan, Enhancing textbooks with visuals from the web for improved learning, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 11931–11944.

[5] J.-H. Yang, C. Lassance, R. Sampaio De Rezende, K. Srinivasan, M. Redi, S. Clinchant, J. Lin, AToMiC: An image/text retrieval test collection to support multimedia content creation, in: Proceedings of the 46th International ACM SIGIR conference on research and development in information retrieval, 2023, pp. 2975–2984.

[6] G. Faggioli, L. Dietz, C. L. Clarke, G. Demartini, M. Hagen, C. Hauff, N. Kando, E. Kanoulas, M. Potthast, B. Stein, et al., Perspectives on large language models for relevance judgment,

[4]The cost of using GPT-4V API for the experiments is around USD 150.

in: Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval, 2023, pp. 39–50.

[7] J. Hessel, A. Holtzman, M. Forbes, R. Le Bras, Y. Choi, CLIPScore: A reference-free evaluation metric for image captioning, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, 2021, pp. 7514–7528.

[8] D. H. Park, S. Azadi, X. Liu, T. Darrell, A. Rohrbach, Benchmark for compositional text-to-image synthesis, in: Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 1), 2021.

[9] J.-H. Kim, Y. Kim, J. Lee, K. M. Yoo, S.-W. Lee, Mutual information divergence: A unified metric for multimodal generative models, in: Advances in Neural Information Processing Systems, volume 35, 2022, pp. 35072–35086.

[10] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, K. Aberman, Dreambooth: Fine tuning text-to-image diffusion models for subject-driven generation, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 22500–22510.

[11] E. Kreiss*, E. Zelikman*, C. Potts, N. Haber, ContextRef: Evaluating referenceless metrics for image description generation, arXiv preprint arxiv:2309.11710 (2023).

[12] 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, 2021, pp. 8748–8763.

[13] H. Liu, C. Li, Q. Wu, Y. J. Lee, Visual instruction tuning, in: Advances in Neural Information Processing Systems, 2023.

[14] H. Liu, C. Li, Y. Li, Y. J. Lee, Improved baselines with visual instruction tuning, in: NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following, 2023.

[15] OpenAI, GPT-4V(ision) system card, 2023.

[16] D. M. Chan, S. Petryk, J. E. Gonzalez, T. Darrell, J. Canny, CLAIR: Evaluating image captions with large language models, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 13638–13646.

[17] Y. Hu, B. Liu, J. Kasai, Y. Wang, M. Ostendorf, R. Krishna, N. A. Smith, TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering, in: 2023 IEEE/CVF International Conference on Computer Vision, 2023, pp. 20349–20360.

[18] D. Chan, A. Myers, S. Vijayanarasimhan, D. Ross, J. Canny, IC3: Image captioning by committee consensus, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 8975–9003.

[19] M. Yuksekgonul, F. Bianchi, P. Kalluri, D. Jurafsky, J. Zou, When and why vision-language models behave like bags-of-words, and what to do about it?, in: The Eleventh International Conference on Learning Representations, 2023.

[20] Y. Lu, X. Yang, X. Li, X. E. Wang, W. Y. Wang, LLMScore: Unveiling the power of large language models in text-to-image synthesis evaluation, arXiv preprint arXiv:2305.11116 (2023).

[21] F. Betti, J. Staiano, L. Baraldi, R. Cucchiara, N. Sebe, Let's ViCE! Mimicking human cognitive behavior in image generation evaluation, in: Proceedings of the 31st ACM International Conference on Multimedia, 2023, p. 9306–9312.

[22] A. Zeng, M. Attarian, brian ichter, K. M. Choromanski, A. Wong, S. Welker, F. Tombari,

A. Purohit, M. S. Ryoo, V. Sindhwani, J. Lee, V. Vanhoucke, P. Florence, Socratic models: Composing zero-shot multimodal reasoning with language, in: The Eleventh International Conference on Learning Representations, 2023.

[23] L. Dietz, J. Dalton, Humans optional? automatic large-scale test collections for entity, passage, and entity-passage retrieval, Datenbank-Spektrum 20 (2020) 17–28.

[24] S. MacAvaney, L. Soldaini, One-shot labeling for automatic relevance estimation, in: Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, 2023, p. 2230–2235.

[25] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, et al., Language models are few-shot learners, in: Advances in Neural Information Processing Systems, volume 33, 2020, pp. 1877–1901.

[26] W. Sun, L. Yan, X. Ma, S. Wang, P. Ren, Z. Chen, D. Yin, Z. Ren, Is ChatGPT good at search? investigating large language models as re-ranking agents, in: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, 2023, pp. 14918–14937.

[27] Z. Qin, R. Jagerman, K. Hui, H. Zhuang, J. Wu, J. Shen, T. Liu, J. Liu, D. Metzler, X. Wang, et al., Large language models are effective text rankers with pairwise ranking prompting, arXiv preprint arXiv:2306.17563 (2023).

[28] E. M. Voorhees, Variations in relevance judgments and the measurement of retrieval effectiveness, in: Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval, 1998, pp. 315–323.

[29] Y. Liu, D. Iter, Y. Xu, S. Wang, R. Xu, C. Zhu, GPTEval: NLG evaluation using GPT-4 with better human alignment, arXiv preprint arXiv:2303.16634 (2023).

[30] N. Pangakis, S. Wolken, N. Fasching, Automated annotation with generative ai requires validation, arXiv preprint arXiv:2306.00176 (2023).

[31] S. Dai, Y. Zhou, L. Pang, W. Liu, X. Hu, Y. Liu, X. Zhang, J. Xu, LLMs may dominate information access: Neural retrievers are biased towards llm-generated texts, arXiv preprint arXiv:2310.20501 (2023).

[32] H. Zhuang, Z. Qin, K. Hui, J. Wu, L. Yan, X. Wang, M. Berdersky, Beyond yes and no: Improving zero-shot LLM rankers via scoring fine-grained relevance labels, arXiv preprint arXiv:2310.14122 (2023).