

Supplementary Information

Predicting semantic consistency between images and Korean hashtags on Instagram

Jiyeon Oh^{1,2} and Jangmin Oh^{1,*}

¹School of AI Convergence, Sungshin Women’s University, Seoul, Republic of South Korea

²College of Business Administration, Seoul National University, Seoul, Republic of South Korea

*Corresponding author: jangmin.oh@sungshin.ac.kr

This document contains the supplementary tables and notes referenced in the main text: Supplementary Tables S1 and S2, Supplementary Note 1 (extended related work) and Supplementary Note 2 (hashtag-level misclassification rates).

Table S1: Per-domain coverage of the dataset (8 domains $\times$ 4 hashtags, 32 in total)

Domain	Keywords	Posts	Pairs	Tags [*]
Travel/Place	#여행 (travel), #여행스타그램 (travel-gram), #혼여행 (solo travel), #국내여행 (domestic travel)	191	1,261	784
Food/Restaurant	#맛집 (foodie spot), #맛스타그램 (tastegram), #먹스타그램 (eatstagram), #JMT (super-tasty)	230	1,540	1,096
Beauty/Shopping	#올리브영 (Olive Young), #메이크업 (makeup), #데일리메이크업 (daily makeup), #꾸안꾸 (effortless styling)	209	1,421	819
Fashion	#패션 (fashion), #데일리룩 (daily look), #오오티디 (OOTD), #패피 (fashionista)	148	1,409	769
Season/Nature	#가을 (autumn), #능소화 (trumpet creeper), #단풍놀이 (foliage outing), #벚꽃 (cherry blossom)	227	1,387	915
Health/Diet	#제로칼로리 (zero calorie), #다이어트식단 (diet meal), #오운완 (workout done), #헬스타그램 (healthstagram)	216	1,911	927
Tech/Product	#ai (AI), #iphone16 (iPhone 16), #갤럭시 (Galaxy), #애플 (Apple)	222	1,431	885
Entertainment	#영화 (movie), #넷플릭스 (Netflix), #스포티파이 (Spotify), #kpop (K-pop)	230	1,651	1,114
Total	32 hashtags	1,673	12,011	6,639

^{*} Unique-tag counts do not sum across domains because some hashtags appear in multiple domains; the total reflects globally distinct hashtag strings.

Table S2: Distribution of consistency labels over the 12,011 annotated image–hashtag pairs.

Consistency score	1	2	3	4	5	Total
Count	1,355	2,272	1,777	2,818	3,789	12,011
Proportion (%)	11.28	18.92	14.79	23.46	31.55	100.00

Supplementary Note 1: Extended related work

This note provides the extended literature review summarized in the Related work section of the main text.

S1.1 Vision–language models and image–text alignment

Recent advances in vision–language models have significantly improved the joint representation of visual and textual modalities. Models such as CLIP and its variants learn aligned image–text embeddings from large-scale web data [1–9] and have demonstrated strong performance on tasks involving image–caption retrieval and zero-shot image classification [10]. These approaches are particularly effective when textual descriptions explicitly refer to visible objects or scenes within an image.

Our adoption of a CLIP-based backbone is motivated by the structural characteristics of this family of models. Some conventional or modular multi-modal approaches combine independently pretrained image and text encoders and fuse their features at a later stage; in this case, the representations of the two modalities reside in different spaces, making it difficult to directly compare the similarity between an image and its text. In contrast, CLIP is trained via contrastive learning on large-scale image–text pairs so that the outputs of each encoder, after projection, are aligned into a comparable joint embedding space. As a result, the cosine similarity between an image embedding and a text embedding provides a relative alignment signal induced by the pretraining objective. As our task is to predict the semantic consistency between an image and a Korean hashtag, the aligned structure of CLIP-based models, which allows the two modalities to be compared in a common space, is well suited as a starting point for this problem.

CLIP-based models, however, exhibit two distinct limitations. First, their pretraining is primarily optimized for global image–text matching, and has been reported to be limited in fine-grained alignment between object parts, regions, or detailed attributes [11–13]. Second, and more directly relevant to the present study, the text used in CLIP pretraining generally consists of image-associated web text, which tends to provide stronger object- or scene-level grounding than affective social-media hashtags. Consequently, when the text deviates from objective description and instead encodes emotion, pragmatics, slang, or cultural context, the reliability of the learned alignment decreases. Whereas the former is a problem of fine-grained visual alignment, the latter concerns relating non-literal, affective meaning to visual information; it is the latter on which this study focuses.

Language-specific adaptations such as KoCLIP have been developed to address linguistic differences [14–17], but they still rely on the fundamental CLIP objective that prioritizes descriptive alignment. This mismatch suggests that performance on conventional benchmarks may not directly translate to real-world social media scenarios, where user-generated text often diverges from literal visual content [10].

S1.2 Hashtag semantics and social-media image–text relations

Hashtags have been widely studied as tools for organizing content, improving searchability, analyzing trends, and understanding user behavior on social-media platforms. Prior work has performed hashtag recommendation, hashtag relevance estimation, and topic inference using textual co-occurrence patterns, visual features, and contextual information surrounding a post [18–22]. These approaches generally regard hashtags as semantic or topical labels that can be inferred from an image or its surrounding textual context.

In real-world social-media environments, however, hashtags serve a variety of communicative purposes beyond the literal description of post content. Users may employ hashtags to express affective states, humor, irony, stance, social identity, community participation, or self-presentation [23–25]. Moreover, some hashtags may be chosen to broaden a post’s exposure or

to participate in a particular trend, even when they are not directly related to the visual content of the image. Consequently, the same hashtag may function as a label that directly describes the visual content in one post, while serving as an affective or pragmatic expression, or as a keyword for exposure, in another.

Previous studies have similarly reported that user-generated hashtags are not always directly grounded in objects or scenes visible in an image. Some hashtags may be only partially related to the image content, depend on contextual information outside the image, or even diverge from its literal meaning [26–28]. These characteristics can make the relationship between an image and a hashtag difficult to determine from visual information alone. Additional information, such as the full caption, co-occurring hashtags, cultural conventions, and the user’s communicative intent, may influence how the hashtag and its consistency with the image are interpreted.

These observations indicate that image–hashtag consistency cannot be reduced to object-level correspondence or general visual relevance alone. The relationship between an image and a hashtag may involve not only visual meaning but also affective, pragmatic, and culturally contextualized interpretation. As a result, approaches that treat hashtags as fixed semantic labels, or that evaluate the image–hashtag relationship as a simple binary distinction between relevant and irrelevant, may fail to capture the varying degrees and forms of consistency observed in real-world social media. This difficulty is particularly pronounced in non-English social media such as Korean, where slang, memes, and culturally specific expressions are frequent and often require background knowledge of the corresponding linguistic and cultural community to assess their consistency with an image.

S1.3 Graded and ordinal multimodal consistency modeling

Several studies have attempted to quantify the consistency or relevance between visual and textual modalities. These studies typically compute similarity scores between image and text embeddings or train classifiers to predict relevance labels for image–text pairs [29–31]. Such approaches are effective at distinguishing relevant from irrelevant pairs in relatively well-defined benchmarks or controlled experimental settings. Studies that identify inconsistencies between images and hashtags as a binary decision can also be viewed as an extension of this relevance-modeling framework [19, 32].

However, existing approaches generally treat the relationship between an image and text as a binary distinction between relevance and irrelevance. In real-world social-media environments, the relationship between the two modalities may instead lie on a continuous spectrum involving different degrees of semantic association. Such graded consistency has received comparatively less attention than binary relevance judgment [25, 33]. More recently, multidimensional perspectives have been proposed to analyze social-media image–text consistency across semantic, affective, and informational dimensions [34]. These perspectives suggest that image–text consistency is not a single binary property, but a complex construct involving multiple levels and dimensions.

Nevertheless, even studies that consider graded consistency have only limitedly modeled or evaluated the ordinal structure among consistency levels. An ordinal structure assumes that adjacent levels are more similar to one another than levels located farther apart on the scale. Accounting for this structure makes it possible to distinguish a one-level prediction error from a large-distance error across opposite ends of the scale. By contrast, evaluations based solely on nominal classification or a single similarity score may not adequately reflect differences in error severity.

This issue is particularly important for affective expressions, metaphors, slang, and culturally contextualized hashtags. The interpretation of such expressions often depends more on users’ affective or pragmatic judgments than on objects directly visible in an image, making consistency judgments inherently subjective and context dependent [28]. To address this gap, we formulate semantic consistency between images and Korean hashtags as a five-level ordinal classification

task. This formulation allows us to evaluate not only whether a prediction exactly matches the reference label, but also the magnitude of prediction errors and the ordinal structure of the task.

S1.4 LLM-assisted annotation and reliability validation

Large language models have recently attracted attention as scalable annotation tools capable of applying consistent labeling criteria to large datasets. In several text-classification tasks, LLMs have been reported to achieve competitive accuracy relative to crowd workers while reducing annotation time and cost [35]. Their use has also expanded to a wide range of tasks, including relevance, stance, sentiment, topic, and semantic-relation annotation, and their broader potential for data annotation and synthesis has been actively discussed [36]. These characteristics make LLM-assisted annotation a practical alternative when assigning fine-grained consistency labels to large numbers of image–hashtag pairs would be difficult through manual annotation alone.

However, the reliability of LLM-generated labels cannot be assumed solely from the general capabilities of the underlying model. Annotation outcomes may vary depending on prompt construction, the wording of evaluation criteria, the examples provided, and the required output format. In tasks with ambiguous decision boundaries, such as semantic proximity assessment, directly transferring human annotation guidelines to an LLM may be insufficient, and even minor prompt modifications can influence the resulting labels [37]. Moreover, the same model may produce inconsistent judgments across repeated evaluations or alternative scoring scales. These findings highlight the need for an explicit rubric, constrained output formats, and separate consistency checks [38].

These concerns are particularly important for subjective annotation tasks in which a single correct answer cannot be clearly established. Judgments concerning affect, stance, or semantic relevance may vary according to annotators’ experiences and interpretations. Therefore, exact agreement between LLM labels and a single aggregated human label may not fully capture annotation quality. In subjective tasks, the distribution of annotator judgments and the disagreement among annotators may themselves contain meaningful information, which can be lost when diverse perspectives are reduced to a single ground-truth label [39]. Human-in-the-loop annotation, in which annotators review and revise model-generated suggestions, may reduce validation costs, but it can also introduce an anchoring effect in which the model’s initial prediction influences subsequent human judgments [40]. Human validation should therefore be designed as an independent comparison rather than as a simple post hoc approval process.

Much of the existing research on LLM-assisted annotation has focused on English-language text tasks, and its findings cannot necessarily be assumed to transfer directly to non-English languages such as Korean or to multimodal judgment tasks. The reliability and consistency of LLM evaluators may vary across languages [41]. In addition, multimodal evaluation requires the joint interpretation of images and text, introducing further challenges related to fine-grained visual evidence, cultural context, and non-literal expressions [42]. Accordingly, this study uses GPT-4o-mini-based automated annotation for large-scale dataset construction while explicitly documenting the annotation prompt and rubric and validating the generated labels against judgments from multiple independent human annotators. We further examine not only exact agreement but also disagreements between adjacent categories and ordinal agreement, thereby providing a more appropriate interpretation of uncertainty in subjective five-level semantic consistency labels.

Supplementary Note 2: Hashtag-level misclassification rates

The following analysis complements the Error structure subsection of the main text.

This tendency is also confirmed quantitatively at the level of individual hashtags. The hashtags #메이크업 (makeup) and #단풍 (autumn foliage), which appear as concrete visual objects

in the image, showed misclassification rates of only 5% and 13%, respectively. In contrast, hashtags whose meaning is not directly visualized showed markedly higher rates. The cultural slang #JMT (a Korean expression meaning “very delicious”) showed a rate of 88%: even when the image clearly depicts food-related content, the model struggles to capture culturally grounded slang that implies taste or quality rather than an explicit visual object, reflecting a *cultural context gap*. Similarly, abstract expressions that convey mood or emotion, such as #여름 (summer, denoting an abstract seasonal mood; 86%), are difficult to align with visual features, reflecting a *metaphorical gap* between affective language and concrete visual entities. In addition, #참교욱, which refers to the title of a specific Korean drama, showed a high rate of 74%, as its consistency is difficult to judge from the image alone without external knowledge of the corresponding content. These results show that the model’s errors are not randomly distributed but vary systematically according to how directly and concretely the meaning of a hashtag is visualized in the image, and whether it requires cultural or external knowledge.

References

- [1] Radford, A. *et al.* Learning transferable visual models from natural language supervision. *arXiv:2103.00020* (2021).
- [2] Jia, C. *et al.* Scaling up visual and vision–language representation learning with noisy text supervision. in *Proceedings of the International Conference on Machine Learning*, 4904–4916 (2021).
- [3] Frome, A. *et al.* Devise: A deep visual-semantic embedding model. in *Advances in Neural Information Processing Systems*, Vol. 26 (2013). URL https://proceedings.neurips.cc/paper_files/paper/2013/file/7cce53cf90577442771720a370c3c723-Paper.pdf.
- [4] Faghri, F., Fleet, D. J., Kiros, J. R. & Fidler, S. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv:1707.05612* (2018). URL <https://arxiv.org/abs/1707.05612>.
- [5] Desai, K. & Johnson, J. *Virtex: Learning visual representations from textual annotations*. in *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 11157–11168 (2021).
- [6] Wang, Z. *et al.* Improving zero-shot generalization for clip with synthesized prompts. *arXiv:2307.07397* (2023). URL <https://arxiv.org/abs/2307.07397>.
- [7] Chen, Z., Liu, G., Zhang, B.-W., Yang, Q. & Wu, L. Altclip: Altering the language encoder in clip for extended language capabilities. in *Findings of the Association for Computational Linguistics: ACL 2023*, 8666–8682 (2023). URL <https://aclanthology.org/2023.findings-acl.552>.
- [8] Chen, Y.-C. *et al.* Uniter: Universal image-text representation learning. in *Computer Vision – ECCV 2020*, 104–120 (Springer-Verlag, 2020).
- [9] Kim, W., Son, B. & Kim, I. Vilt: Vision-and-language transformer without convolution or region supervision. in *International conference on machine learning*, 5583–5594 (PMLR, 2021).
- [10] Gwinnup, J. & Duh, K. A survey of vision-language pre-training from the lens of multimodal machine translation. *arXiv:2306.07198* (2023). URL <https://arxiv.org/abs/2306.07198>.

- [11] Asokan, M., Wu, K. & Albreiki, F. Finelip: Extending clip’s reach via fine-grained alignment with longer text inputs. *arXiv:2504.01916* (2025). URL <https://arxiv.org/abs/2504.01916>.
- [12] Xie, C. *et al.* Fg-clip: Fine-grained visual and textual alignment. *arXiv:2505.05071* (2025). URL <https://arxiv.org/abs/2505.05071>.
- [13] Liu, J., Zhou, H., Liu, Y., Chen, B. & Jiang, Y. Cross modal fine-grained alignment via granularity-aware and region-uncertain modeling. *arXiv:2511.07710* (2025). URL <https://arxiv.org/abs/2511.07710>.
- [14] jaketae. Koclip: Korean clip (contrastive language-image pretraining). <https://github.com/jaketae/koclip> (2024). Accessed: 2025-11-01.
- [15] Ma, Y. *et al.* X-clip: End-to-end multi-grained contrastive learning for video-text retrieval. in *Proceedings of the 30th ACM international conference on multimedia*, 638–647 (2022).
- [16] Takahashi, R. & Asano, Y. M. Release of pre-trained models for the japanese language. *arXiv:2404.01657* (2024). URL <https://arxiv.org/abs/2404.01657>.
- [17] Wang, J., Wang, H., Deng, J., Wu, W. & Zhang, D. Efficientclip: Efficient cross-modal pre-training by ensemble confident learning and language modeling. *arXiv:2109.04699* (2021).
- [18] Zhang, S. *et al.* Hashtag recommendation for photo sharing services. *Proceedings of the AAAI Conference on Artificial Intelligence* **33**, 5805–5812 (2019). URL <https://ojs.aaai.org/index.php/AAAI/article/view/4528>.
- [19] Dadgar, S. & Neshat, M. A novel hybrid multi-modal deep learning for detecting hashtag incongruity on social media. *Sensors* **22**, 9870 (2022).
- [20] Qadir, A. & Riloff, E. *Learning emotion indicators from tweets: Hashtags, hashtag patterns, and phrases*. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (2014).
- [21] Tan, C. S. N. A late fusion multimodal sentiment analysis model and hashtags summarizer with interactive visualisation for geotagged instagram posts (2019). URL <https://dr.ntu.edu.sg/handle/10356/77119>. Thesis, Nanyang Technological University, Singapore.
- [22] Goel, P. & Thareja, R. *Emotion analysis of twitter data using hashtag emotions*. in *Applications of Computing and Communication Technologies*, 88–98 (Springer Singapore, 2018).
- [23] Felbo, B., Mislove, A., Søgaard, A., Rahwan, I. & Lehmann, S. *Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm*. in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 1615–1625 (2017).
- [24] You, Q., Luo, J., Jin, H. & Yang, J. *Robust image sentiment analysis using progressively trained and domain transferred deep networks*. in *Proceedings of the AAAI conference on Artificial Intelligence*, Vol. 29 (2015).
- [25] Das, R. & Singh, T. D. Multimodal sentiment analysis: A survey of methods, trends, and challenges. *ACM Computing Surveys* **55** (2023). URL <https://doi.org/10.1145/3586075>.
- [26] Giannoulakis, S. & Tsapatsoulis, N. Evaluating the descriptive power of instagram hash-tags. *Journal of Innovation in Digital Ecosystems* **3**, 114–129 (2016). URL <https://www.sciencedirect.com/science/article/pii/S2352664516300141>.

- [27] Giannoulakis, S. & Tsapatsoulis, N. Filtering instagram hashtags through crowdtagging and the hits algorithm. *IEEE Transactions on Computational Social Systems* **6**, 592–603 (2019).
- [28] Kassem, G., Asfoura, E., Alhuthaifi, B., Gallego, J. & Belhaj, F. Sentiment analysis and classifying hashtags in social media using data mining techniques. *Information Sciences Letters* **12**, 2153–2163 (2023).
- [29] Xiao, R., Kim, S., Georgescu, M.-I., Akata, Z. & Alaniz, S. *FLAIR: VLM with fine-grained language-informed image representations*. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* (2025).
- [30] Gomes, G., Zerva, C. & Martins, B. Evaluation of multilingual image captioning: How far can we get with clip models? *arXiv:2502.06600* (2025). URL <https://arxiv.org/abs/2502.06600>.
- [31] Ye, K., Lu, Z., Shen, L. & Cui, T. Rewarding fine-grained image captioning with keyword group contrastive. *Expert Systems with Applications* 131405 (2026). URL <https://www.sciencedirect.com/science/article/pii/S0957417426003180>.
- [32] Ha, Y., Park, K., Joo, J. & Cha, M. Automatically detecting image–text mismatch on instagram with deep learning. *Journal of Advertising* **50**, 52–62 (2021). URL <https://doi.org/10.1080/00913367.2020.1843091>.
- [33] Chandrasekaran, G., Nguyen, T. N. & Hemanth D., J. Multimodal sentimental analysis for social media applications: A comprehensive review. *WIREs Data Mining and Knowledge Discovery* **11**, e1415 (2021). URL <https://wires.onlinelibrary.wiley.com/doi/abs/10.1002/widm.1415>.
- [34] Xia, H. *et al.* A multidimensional framework for analyzing image–text consistency in social media. *Applied Sciences* (2026). URL <https://api.semanticscholar.org/CorpusID:287680322>.
- [35] Gilardi, F., Alizadeh, M. & Kubli, M. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences* **120**, e2305016120 (2023). URL <https://doi.org/10.1073/pnas.2305016120>.
- [36] Tan, Z. *et al.* *Large language models for data annotation and synthesis: A survey*. in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 930–957 (Association for Computational Linguistics, Miami, Florida, USA, 2024). URL <https://aclanthology.org/2024.emnlp-main.54/>.
- [37] Yadav, S., Chopra, T. & Schlechtweg, D. Towards automating text annotation: A case study on semantic proximity annotation using GPT-4 (2024). URL <https://arxiv.org/abs/2407.04130>. arXiv:2407.04130.
- [38] Lee, N., Hong, J. & Thorne, J. *Evaluating the consistency of LLM evaluators*. in *Proceedings of the 31st International Conference on Computational Linguistics*, 10650–10659 (Association for Computational Linguistics, Abu Dhabi, UAE, 2025). URL <https://aclanthology.org/2025.coling-main.710/>.
- [39] Pavlovic, M. & Poesio, M. *The effectiveness of LLMs as annotators: A comparative overview and empirical analysis of direct representation*. in *Proceedings of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @ LREC-COLING 2024*, 100–110 (ELRA and ICCL, Torino, Italia, 2024). URL <https://aclanthology.org/2024.nlperspectives-1.11/>.

- [40] Schroeder, H., Roy, D. & Kabbara, J. *Just put a human in the loop? investigating LLM-assisted annotation for subjective tasks.* in *Findings of the Association for Computational Linguistics: ACL 2025*, 25771–25795 (Association for Computational Linguistics, Vienna, Austria, 2025). URL <https://aclanthology.org/2025.findings-acl.1323/>.
- [41] Fu, X. & Liu, W. *How reliable is multilingual LLM-as-a-judge?* in *Findings of the Association for Computational Linguistics: EMNLP 2025*, 11040–11053 (Association for Computational Linguistics, Suzhou, China, 2025). URL <https://aclanthology.org/2025.findings-emnlp.587/>.
- [42] Zhang, X. *et al.* GPT-4V(ision) as a generalist evaluator for vision–language tasks (2023). URL <https://arxiv.org/abs/2311.01361>. arXiv:2311.01361.