

Predicting semantic consistency between images and Korean hashtags on Instagram

Jiyeon Oh

Seoul National University

Jangmin Oh

`jangmin.oh@sungshin.ac.kr`

Sungshin Women's University

Article

Keywords:

Posted Date: October 22nd, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10982008/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Predicting semantic consistency between images and Korean hashtags on Instagram

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

^{1*}School 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(s). E-mail(s): jangmin.oh@sungshin.ac.kr;
Contributing authors: jynoh825@snu.ac.kr;

Abstract

On image-centric social media, hashtags convey emotion, slang, and cultural context rather than literal descriptions, so their consistency with the image is graded and ambiguous—properties that descriptively trained vision–language models capture poorly. We formulate image–Korean–hashtag consistency as a five-level ordinal classification task and compare Contrastive Language–Image Pre-training (CLIP) and its Korean-adapted variant KoCLIP under three strategies: similarity-based prediction, frozen-backbone classification, and end-to-end fine-tuning. Supervision comes from 12,011 image–hashtag pairs from public Korean Instagram posts annotated with GPT-4o mini under an explicit rubric and validated against four independent human annotators with near-human ordinal agreement (mean quadratic-weighted $\kappa = \mathbf{0.68}$). Under post-level grouped five-fold cross-validation, fine-tuned KoCLIP performed best (accuracy 45.8%, macro F1 0.439), whereas the same protocol degraded CLIP below its frozen counterpart: fine-tuning benefits are gated by the linguistic suitability of the pretrained representation. Classical classifiers on frozen KoCLIP embeddings remained competitive on aggregate metrics, but fine-tuned KoCLIP kept a statistically significant class-balanced (macro F1) advantage over the strongest, extreme gradient boosting (XGBoost). Since human annotators agreed exactly on only 52.4% of validated pairs, much residual error reflects intrinsic task ambiguity. These results argue for graded formulation, ordinal-aware evaluation, and validated automated annotation in affect-aware multimodal understanding of non-English social media.

Keywords: Instagram, hashtags, multimodal learning, semantic consistency, CLIP, KoCLIP

1 Introduction

On image-centric social-media platforms such as Instagram, hashtags function as more than indexing metadata; they serve as a means of user expression that conveys emotions, contextual information, and cultural nuances [1]. Unlike conventional image labels used in computer-vision tasks, such as “dog” or “cat,” user-generated hashtags frequently extend beyond literal descriptions of visual content. For example, a photo of a coffee cup may be tagged with abstract or affective expressions such as #힐링 (healing) or #월요병 (Monday blues) rather than labels denoting concrete objects. This discrepancy poses a substantial challenge for quantitatively evaluating the “semantic consistency” between an image and its text, which is closely tied to content retrieval and recommendation, post understanding, and user behavior analysis [2].

In real-world social media, hashtags not only describe the content of an image but also serve a variety of communicative functions, such as expressing emotion, humor, irony, encouragement, self-presentation, and community participation [2, 3]. Korean social-media content, in particular, frequently includes abbreviations, slang, memes, and culturally specific expressions. Such expressions may not correspond directly to identifiable visual entities and may be only indirectly related to, or even conflict with, the literal content of an image. Consequently, annotators may assign different levels of consistency to the same image–hashtag pair depending on their experiences, interpretations, and understanding of the relevant cultural context.

This semantic ambiguity distinguishes image–hashtag consistency prediction from conventional object-level image–text alignment. When text directly describes an object or scene, the visual and linguistic modalities share relatively explicit referents. Interpreting a social-media hashtag, however, may require consideration of affective meaning, implicit context, pragmatic intent, and socially or culturally grounded language in addition to information directly visible in the image [4]. Moreover, the relationship between an image and a hashtag may not be adequately represented as a binary distinction between relevance and irrelevance. Instead, it may lie on a graded spectrum ranging from highly inconsistent to highly consistent. This property complicates the interpretation of exact accuracy against a single reference label and introduces uncertainty, particularly when distinguishing between adjacent consistency levels.

Recent advances in vision–language representation learning, exemplified by CLIP and its variants, have demonstrated strong performance by aligning images and text within a comparable embedding space [5]. Although CLIP-based models are trained on diverse web-based image–text data, their contrastive learning objective primarily rewards overall correspondence between images and their associated text. These models are therefore effective for expressions grounded in visible objects, attributes, and scenes, but may be less capable of representing non-literal, metaphorical, affective,

and culturally contextualized hashtags [6]. A Korean-adapted model such as KoCLIP may reduce the linguistic mismatch associated with Korean input, but language adaptation alone does not guarantee that pretrained embedding similarity will correspond directly to subjective, fine-grained consistency levels [7].

This study systematically analyzes the extent to which existing vision-language models can capture subjective semantic consistency between images and Korean Instagram hashtags. We formulate image-hashtag consistency as a five-level ordinal classification task, with labels ranging from 1 (highly inconsistent) to 5 (highly consistent). We then compare a general-purpose CLIP backbone with the Korean-adapted KoCLIP under three prediction and training strategies: similarity-based prediction, frozen-backbone classification, and end-to-end fine-tuning. This design enables us to separately examine the effects of linguistic suitability in the pretrained backbone, a supervised decision boundary, and task-specific representation adaptation.

Assigning fine-grained labels to a large number of subjective image-hashtag pairs requires substantial time and resources. We therefore construct the dataset using a GPT-4o-mini-based automated annotation procedure with an explicit annotation rubric and constrained output format, building on recent developments in LLM-assisted data annotation [8, 9]. However, we do not treat the automatically generated labels as definitive ground truth. Instead, we conduct an independent multi-annotator human validation study and examine exact agreement, disagreement between adjacent categories, and ordinal agreement. This approach is intended to retain the scalability of automated annotation while explicitly evaluating uncertainty in subjective consistency judgments.

Our experiments yield three main findings. First, cosine similarity between pretrained embeddings alone is insufficient to reliably distinguish the five consistency levels, whereas a trainable classifier over frozen multimodal representations improves class discrimination and class-balanced performance for both backbones. Second, end-to-end fine-tuning is not beneficial in itself: under an identical training protocol, fine-tuned KoCLIP achieved the strongest and most balanced class-wise performance among the six main configurations, whereas fine-tuning CLIP degraded performance below its frozen-backbone counterpart. The benefit of adaptation is thus gated by the linguistic suitability of the pretrained representation. Third, classical classifiers such as SVM and XGBoost trained on frozen KoCLIP embeddings remain competitive with fine-tuned KoCLIP on aggregate metrics—a post-level cluster bootstrap detects no significant difference between fine-tuned KoCLIP and the strongest of them, XGBoost, in accuracy or quadratic weighted kappa—while the macro-F1 advantage of fine-tuned KoCLIP over XGBoost is statistically significant. The principal contribution of full fine-tuning therefore lies not in raising aggregate accuracy but in stabilizing performance across the imbalanced consistency levels, including the low-consistency pairs that downstream filtering applications must detect. These findings should also be read against the subjectivity of the task itself: even human annotators agreed exactly on only 52.4% of the validated pairs, with most disagreements confined to adjacent consistency levels.

The main contributions of this study are summarized as follows:

- We formulate image–Korean–hashtag consistency as a five-level ordinal prediction task rather than binary relevance or a fixed similarity score, capturing both the subjectivity of social-media language and the severity of prediction errors.
- We construct a 12,011-pair dataset from public Korean Instagram posts by combining rubric-guided GPT-4o mini annotation with independent multi-annotator human validation, enabling analysis of label reliability and annotation uncertainty.
- We systematically compare CLIP and KoCLIP under similarity-based prediction, frozen-backbone classification, and end-to-end fine-tuning, separating the effects of backbone language adaptation, supervised decision boundaries, and representation adaptation.
- We evaluate with post-level grouped cross-validation, nominal and ordinal metrics, and post-level cluster-bootstrap and McNemar tests, and use SVM and XGBoost ablations and a failure analysis to show that the significant advantage of full fine-tuning is concentrated in class-balanced performance.

2 Related work

2.1 Overview of prior work

An extended review of the literature is provided in Supplementary Note 1; here we summarize the four strands most relevant to our design. First, contrastively trained vision–language models such as CLIP and its variants [5, 6] align images and text in a shared embedding space, which is why we adopt them as backbones; however, their pretraining rewards global, descriptive image–text matching, so the learned alignment weakens for text that encodes affect, pragmatics, slang, or cultural context, and language-specific adaptations such as KoCLIP [7] inherit this descriptive objective. Second, studies of hashtag semantics show that hashtags serve communicative functions beyond describing content—emotion, humor, stance, self-presentation, and exposure seeking—so that the same hashtag may be a literal label in one post and an affective or pragmatic marker in another, especially in non-English communities where slang and memes are frequent [2–4]. Third, existing work on image–text consistency has mostly treated the relationship as binary relevance; graded consistency has received less attention [10, 11], and even graded formulations rarely model or evaluate the ordinal structure among levels, which is essential for distinguishing a one-level error from an error across opposite ends of the scale. Fourth, large language models have become scalable annotation tools [9, 12], but the reliability of their labels depends on prompt and rubric design and must be validated independently against multiple human annotators rather than assumed [8, 13]; such evidence remains scarce for non-English, multimodal judgments [14, 15].

2.2 Research gap and positioning of this study

Taken together, the preceding discussion reveals several gaps in existing research. First, work on image–text relevance has focused primarily on retrieval, embedding similarity, or binary relevant-versus-irrelevant classification using descriptive captions. In real-world social media, however, hashtags do not merely describe the literal content of an

image but also express emotion, pragmatic intent, trends, and cultural context. Korean
 hashtags in particular frequently include slang, abbreviations, memes, and culturally
 specific expressions, yet the consistency between such expressions and images has
 rarely been analyzed as a graded, ordinal relationship.

Second, it remains insufficiently understood how general-purpose vision–language
 representations and language-adapted representations behave differently on this task.
 Prior work has mostly used the similarity of pretrained embeddings directly, or
 reported the performance of a particular classification setup in isolation; few studies
 have systematically separated and compared the effects of the backbone’s linguis-
 tic suitability and the task-adaptation strategy under the same data and evaluation
 conditions.

Third, a question remains as to how annotation reliability can be ensured
 and evaluated when constructing subjective multimodal consistency labels at scale.
 LLM-assisted annotation can improve the scalability and consistency of dataset con-
 struction, but its results may be affected by prompt design, the scoring rubric, and
 the model’s judgment bias. Moreover, since fine-grained semantic consistency can be
 judged differently even among human annotators, the exact agreement between an
 LLM label and a single human label alone is insufficient to fully characterize the
 reliability of the data.

To address these gaps, our study formulates the semantic consistency between
 an image and a Korean Instagram hashtag as a five-level ordinal classification
 problem, and systematically analyzes the capabilities and limitations of existing mul-
 timodal models—general-purpose CLIP and language-adapted KoCLIP—under the
 same conditions.

3 Methods

This section is structured to address the research questions defined in the Research
 objectives and problem formulation subsection. We first formulate semantic con-
 sistency prediction as a five-level ordinal classification problem and describe the
 construction and annotation of the dataset, including the automated labeling proce-
 dure and the human validation protocol. We then present the model configurations
 (Table 1), which vary the multimodal backbone (CLIP or KoCLIP) and the prediction
 and training strategy (similarity-based prediction, frozen-backbone classification, or
 end-to-end fine-tuning), followed by the experimental setup, the evaluation protocol
 and metrics, the statistical analysis, and the ethics and AI-use statements.

3.1 Research objectives and problem formulation

The primary objective of this study is to empirically investigate how effectively mul-
 timodal models can understand semantic consistency between images and Korean
 Instagram hashtags, and to analyze the extent to which domain-specific fine-tuning
 can mitigate their limitations.

Unlike conventional image–text alignment tasks that assume binary relevance,
 semantic consistency between hashtags and images exists on a continuous and sub-
 jective spectrum. This observation is consistent with prior studies on hashtag usage,

Table 1 Model configurations for semantic consistency prediction

Experiment	Backbone	Prediction Method	Backbone Update
Baseline 1-1	CLIP	Cosine similarity (threshold)	Frozen
Baseline 1-2	KoCLIP	Cosine similarity (threshold)	Frozen
Baseline 2-1	CLIP	Frozen embedding + MLP	Frozen
Baseline 2-2	KoCLIP	Frozen embedding + MLP	Frozen
End-to-End 3-1	CLIP	Full fine-tuning + MLP	Trainable
End-to-End 3-2	KoCLIP	Full fine-tuning + MLP	Trainable

which report that hashtags often encode affective, pragmatic, or contextual meaning rather than literal visual descriptions [2, 4]. To reflect this characteristic, we reformulate the consistency judgment as a five-level ordinal classification problem, where each image–hashtag pair is assigned a discrete consistency score ranging from 1 (highly inconsistent) to 5 (highly consistent).

Based on this formulation, we address the following research questions:

- **RQ1:** How well do similarity-based approaches capture graded semantic consistency between images and Korean hashtags?
- **RQ2:** Does introducing a trainable classification head over multimodal representations improve performance compared with similarity-based prediction?
- **RQ3:** How does end-to-end fine-tuning affect performance relative to frozen-backbone classification, and how does this effect differ between CLIP and KoCLIP?

These questions guide the design of our dataset, model configurations, and experimental evaluation.

3.2 Data collection and construction

To evaluate semantic consistency in real-world social media settings, a dataset consisting of Korean Instagram posts was constructed. However, manual annotation at scale is challenging and prone to inconsistency [3]. Therefore, given the subjective and context-dependent nature of hashtag usage, the data construction process was designed to capture a wide range of topics while ensuring labeling reliability through a multi-stage validation approach.

3.2.1 Domain and hashtag selection

To build a representative dataset, we selected 32 hashtags spanning eight domains that are prevalent on Korean Instagram: Travel/Place, Food, Beauty/Shopping, Fashion, Season/Nature, Health/Diet, Tech/Product, and Entertainment. Rather than simply enumerating popular topics, we chose these domains according to two criteria aligned with the goal of this study. First, we ensured that the linguistic character of the hashtags was diverse: the selected domains yield not only tags that literally describe

the image content (e.g., #국내여행 (domestic travel), #벚꽃 (cherry blossom)), but also affective, connotative, or slang expressions (e.g., #꾸안꾸 (effortless styling), #JMT (super-tasty)), so that the emotional and cultural ambiguity of Korean hashtags that motivates our work is genuinely reflected in the data. Second, we sought to obtain a broad distribution of image–hashtag semantic consistency: by including domains that are visually grounded (e.g., Season/Nature) alongside domains dominated by abstract or context-dependent tags (e.g., Tech/Product, Entertainment), the five-level consistency scale is observed across its full range rather than concentrated at particular values. We assigned four representative hashtags to each of the eight domains (32 in total; Supplementary Table S1), maintaining balance across domains.

3.2.2 Data crawling and preprocessing

The data were collected using a web crawler built with the Python Selenium library. For each of the 32 hashtags, we retrieved the most recent public posts, restricting collection to image posts and excluding video (Reels) content. From each post we extracted the primary image together with its full list of associated hashtags. Captions were cleaned by removing metadata prefixes, URLs, @-mentions, and non-semantic repetition, while emoji were preserved as affective signals. We then paired the primary image of each post with each of its associated hashtags and removed duplicated entries, inaccessible posts, and pairs with missing images or hashtags. The resulting dataset comprises 12,011 valid image–hashtag pairs drawn from 1,673 posts and 6,639 distinct hashtag strings, with an average of 7.2 hashtags per post. This dataset was then annotated using the automated procedure described in the next subsection.

3.2.3 Automated consistency labeling with a large language model

Manually annotating semantic consistency for all 12,011 image–hashtag pairs described above is inherently subjective and difficult to scale. To ensure a consistent labeling standard, we employed the structured-output functionality of OpenAI’s GPT-4o mini (`gpt-4o-mini`) for automated annotation, following a strategy similar to [8]. This automated pass constitutes the first stage of the multi-stage validation approach outlined at the beginning of this section; its reliability is examined in the human validation study described below and reported in Results.

Annotation was carried out at the post level rather than the pair level. For each post, the model received the image, the complete list of hashtags attached to that post (on average 7.2 per post) in a single request, and returned a separate consistency score for every hashtag in that list. The hashtags to be scored were therefore not restricted to the 32 seed hashtags used for collection, but comprised all hashtags actually attached to the post. Presenting the hashtags of a post together allows the annotator to evaluate them under a shared interpretation of the visual content, preventing the assessment of the same image from drifting from one hashtag to the next. A consequence of this design is that labels belonging to the same post are not statistically independent, which directly motivates the post-level grouping adopted in our cross-validation protocol (see the Evaluation protocol subsection).

323 The model was instructed to act as an expert annotator and to assign a consistency
324 score on a five-point Likert scale: 1 (Very Low), 2 (Low), 3 (Medium), 4 (High), and
325 5 (Very High).

326 Model responses were constrained to a typed output schema, which required the
327 model to emit, for each hashtag, an integer score in $\{1, \dots, 5\}$ together with a brief
328 justification and the visual evidence supporting that score. Schema validation elimi-
329 nated malformed and out-of-range outputs and removed the need for post-hoc parsing
330 of free-form text. The annotation guideline instructed the model to (i) give priority
331 to visual evidence in the image, (ii) use the caption only as secondary context, (iii)
332 interpret Korean slang and affective expressions (e.g., #JMT (super-tasty), #꾸안꾸
333 (effortless styling)) according to their idiomatic rather than literal meaning, and (iv)
334 assign a conservative score when the visual evidence was insufficient. Rule (iii) is of par-
335 ticular importance here: the connotative and slang expressions deliberately included
336 during domain selection would otherwise be read literally and their consistency sys-
337 tematically underestimated. Decoding was configured with `temperature` = 0.0 to
338 reduce sampling variability, and failed requests were retried with exponential backoff
339 and logged separately. The complete guideline, including the full prompt text, scoring
340 rubric, boundary-case rules, and the output schema, is provided in the accompanying
341 code repository (see the Data availability statement); the repository’s `README` indicates
342 where each of these materials is located.

343 The distribution of the image–hashtag consistency labels generated with GPT-4o
344 mini (`gpt-4o-mini`) is reported in Supplementary Table S2. Overall, the higher end
345 of the consistency scale accounts for a larger share, with scores of 4 and 5 together
346 making up roughly 55% of all pairs. Score 3 accounts for 14.79% and score 2 for
347 18.92%, while the lowest grade (score 1) is the least frequent at 11.28%.

348

349

350

3.2.4 Human validation of automated labels

351 To evaluate the reliability of the GPT-4o mini annotations, we randomly selected
352 500 image–hashtag pairs from the expanded dataset. Four undergraduate research
353 assistants familiar with Korean social-media language independently annotated all 500
354 pairs using the same five-level scoring rubric employed for GPT-4o mini annotation.
355 Each annotator evaluated the samples independently and did not have access to the
356 GPT-4o-mini-generated labels or the judgments of the other annotators.

357 Before annotation, the annotators were provided with written definitions of the five
358 consistency levels and representative examples illustrating literal, affective, metaphori-
359 cal, and unrelated image–hashtag relationships. No post-hoc discussion or adjudication
360 was performed before the reliability analysis, thereby preserving the independence of
361 the individual judgments.

362

363

364

3.2.5 Annotation reliability measures

365 Because the consistency labels form an ordinal scale, annotation reliability was
366 evaluated using measures that account for the ordered relationship among the five cat-
367 egories. Ordinal Krippendorff’s alpha was used as the primary measure of agreement
368 among the four human annotators. Krippendorff’s alpha accommodates more than

two annotators and allows disagreement to be defined according to the measurement level of the labels [16, 17].

We additionally computed quadratic-weighted Cohen’s kappa for all six human–human annotator pairs and all four GPT-4o mini–human annotator pairs [18]. Quadratic weighting assigns a larger penalty to disagreements involving distant categories than to disagreements involving adjacent categories. Pairwise results were summarized using their mean, standard deviation, and range.

To facilitate intuitive interpretation, we also report exact agreement, adjacent agreement, and mean absolute error. Exact agreement denotes the proportion of samples receiving identical scores, whereas adjacent agreement denotes the proportion for which the two scores differed by at most one level. Mean absolute error measures the average absolute distance between two ratings. A 95% confidence interval for ordinal Krippendorff’s alpha was estimated by resampling the 500 annotation units with replacement 2,000 times.

3.3 Model architecture and configurations

To systematically examine how different modeling choices affect performance, we organize the experimental design along two dimensions: the multimodal backbone and the prediction and training strategy. For the multimodal backbone, we compare CLIP, a general-purpose vision–language model, with KoCLIP, which incorporates Korean-language characteristics. The second dimension concerns how the final consistency class is derived from pretrained multimodal representations and consists of three strategies: similarity-based prediction (Baseline 1), classification over embeddings extracted from a frozen multimodal backbone (Baseline 2), and end-to-end fine-tuning.

This design enables the contribution of each component to be evaluated in a controlled manner. Comparing CLIP and KoCLIP under the same prediction and training strategy isolates the effect of the Korean-adapted backbone. Within the same backbone, comparing similarity-based prediction with frozen-backbone classification evaluates the contribution of a learnable decision boundary. Comparing frozen-backbone classification with end-to-end fine-tuning further assesses the additional effect of adapting the multimodal representations to the target task. The complete set of experimental configurations is summarized in Table 1.

Figure 1 illustrates the overall processing pipeline shared by the learning-based configurations. Each configuration uses a multimodal backbone to generate image and text embeddings, fuses the resulting representations, and passes them to a prediction module.

3.3.1 Shared multimodal representation

As shown in Figure 1, all learning-based configurations share the same encoding stage. An image I and a hashtag h are transformed into embedding vectors by the image encoder and text encoder of the selected multimodal backbone:

$$\mathbf{v} = f_{\text{img}}(I; \theta_{\text{img}}), \quad \mathbf{t} = f_{\text{text}}(h; \theta_{\text{text}}). \quad (1)$$

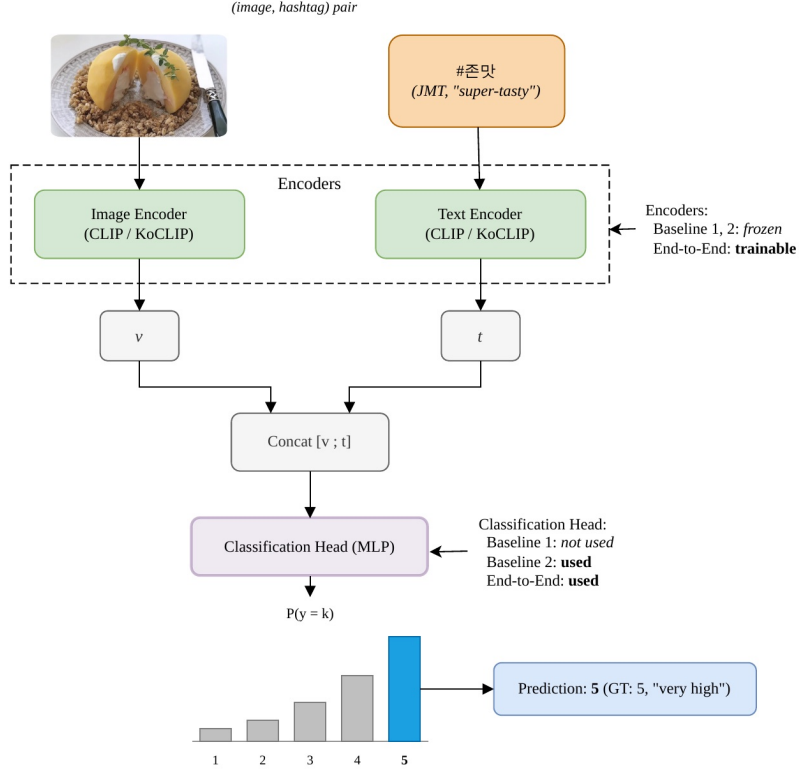


Fig. 1 End-to-end fine-tuning configuration. The key differences across experimental settings are: (i) whether the multimodal encoders are *frozen* (Baselines 1–2) or **trainable** (End-to-End), and (ii) whether a **classification head** is *absent* (Baseline 1) or **present** (Baseline 2 and End-to-End). The pipeline is illustrated with a real image–hashtag example (#존맛 (JMT, “super-tasty”); the photograph was taken by the authors), for which the model outputs a distribution over the five consistency levels and predicts level 5 (ground truth: 5). Photo credit: Jiyeon Oh.

where $\mathbf{v}$ and $\mathbf{t}$ denote the image and text representations, respectively, and θ_{img} and θ_{text} are the parameters of the image and text encoders in the selected CLIP or KoCLIP backbone. The subsequent three configurations are defined according to how the image and text representations are compared or combined, and whether the backbone parameters are frozen or updated.

3.3.2 Baseline 1: similarity-based prediction

Baseline 1 is designed to evaluate whether the pretrained image–text embedding space of CLIP or KoCLIP provides sufficient information for predicting the graded semantic consistency between an image and a Korean hashtag. In this setting, no separate classification head is trained. Instead, the cosine similarity between the image embedding $\mathbf{v}$ and the text embedding $\mathbf{t}$ is computed as

$$s(\mathbf{v}, \mathbf{t}) = \frac{\mathbf{v}^\top \mathbf{t}}{\|\mathbf{v}\|_2 \|\mathbf{t}\|_2}. \quad (2)$$

Here, $s(\mathbf{v}, \mathbf{t})$ denotes the embedding similarity between the image and the hashtag. A higher similarity indicates that the two inputs are located closer to each other in the pretrained multimodal embedding space, whereas a lower similarity indicates weaker semantic alignment.

To convert the continuous similarity score into one of the five consistency classes, four thresholds, $\tau_1, \tau_2, \tau_3, \tau_4$ are used. The predicted class $\hat{y}$ is defined as

$$\hat{y} = \begin{cases} 1, & s < \tau_1, \\ 2, & \tau_1 \leq s < \tau_2, \\ 3, & \tau_2 \leq s < \tau_3, \\ 4, & \tau_3 \leq s < \tau_4, \\ 5, & s \geq \tau_4. \end{cases} \quad (3)$$

For each fold, the thresholds are selected to maximize macro F1 on the corresponding outer training partition. Specifically, candidate threshold sets satisfying the ordering constraint $\tau_1 < \tau_2 < \tau_3 < \tau_4$ are evaluated, and the set yielding the highest macro-F1 on the training partition is selected. The resulting thresholds are then applied to the held-out test partition of that fold, while the test data are not used during threshold selection.

In this configuration, the parameters of the CLIP or KoCLIP backbone remain fixed, and no trainable classifier is introduced. Baseline 1 therefore assesses how well graded semantic consistency can be predicted using only pretrained multimodal embeddings and a threshold-based decision rule calibrated on the training data.

3.3.3 Baseline 2: frozen-backbone classification

Baseline 2 is designed to assess whether introducing a trainable classifier on top of fixed pretrained multimodal representations enables effective prediction of the graded semantic consistency between an image and a Korean hashtag. In this setting, the image-encoder and text-encoder parameters of CLIP or KoCLIP, denoted by θ_{img} and θ_{text} are fully frozen. Only the image representation $\mathbf{v}$ and text representation $\mathbf{t}$ extracted from the two encoders are used as inputs to the classifier.

The two embeddings are concatenated to form a single multimodal representation:

$$\mathbf{z} = [\mathbf{v}; \mathbf{t}]. \quad (4)$$

The fused representation $\mathbf{z}$ is then passed through a shallow multilayer perceptron (MLP) classification head and transformed into logits for the five consistency classes:

$$\mathbf{h} = \text{Dropout}(\text{ReLU}(\mathbf{W}_1 \mathbf{z} + \mathbf{b}_1)), \quad (5)$$

$$\mathbf{o} = \mathbf{W}_2 \mathbf{h} + \mathbf{b}_2, \quad (6)$$

$$p(y = k \mid I, h) = \text{softmax}(\mathbf{o})_k, \quad \hat{y} = \arg \max_{k \in \{1, \dots, 5\}} p(y = k \mid I, h). \quad (7)$$

Here, $\mathbf{h}$ denotes the hidden representation, $\mathbf{o}$ is the logit vector over the five consistency classes, and $\psi = \{\mathbf{W}_1, \mathbf{b}_1, \mathbf{W}_2, \mathbf{b}_2\}$ represents the trainable parameters of the classification head. In our implementation, the concatenated embedding is projected into a 256-dimensional hidden space, followed by a ReLU activation and dropout, and a final linear layer produces the logits for the five classes.

During training, the backbone parameters remain fixed and only the parameters of the classification head are optimized:

$$\theta_{\text{img}}, \theta_{\text{text}} \text{ fixed}, \quad \psi \text{ optimized.} \quad (8)$$

Baseline 2 therefore evaluates whether a trainable non-linear decision boundary can improve semantic consistency prediction over similarity-based prediction without modifying the pretrained CLIP or KoCLIP representations. It also serves as a reference configuration for the subsequent comparison with end-to-end fine-tuning, allowing us to isolate the additional effect of adapting the multimodal representations themselves to the target task.

3.3.4 End-to-end fine-tuning

End-to-End fine-tuning jointly optimizes the pretrained multimodal backbone and the classification head, thereby adapting the image and text representations themselves to the prediction of graded semantic consistency between an image and a Korean hashtag. As in Baseline 2, the image representation $\mathbf{v}$ and text representation $\mathbf{t}$ are concatenated into a single multimodal representation $\mathbf{z}$, which is fed into the MLP classification head. However, unlike Baseline 2, which freezes the parameters of the image and text encoders, end-to-end fine-tuning jointly optimizes the backbone parameters θ_{img} and θ_{text} together with the classification-head parameters ψ :

$$\theta_{\text{img}}, \theta_{\text{text}}, \psi \text{ jointly optimized.} \quad (9)$$

This allows the multimodal representations, originally pretrained for generic image-text alignment, to adapt to the consistency-prediction task, which involves the affective, contextual, and cultural characteristics of Korean hashtags.

The model is trained by minimizing a class-weighted cross-entropy loss to mitigate the effect of class imbalance:

$$\mathcal{L} = -\frac{1}{N} \sum_{i=1}^N \sum_{c=1}^5 w_c y_{ic} \log \hat{p}_{ic}, \quad (10)$$

where N is the number of samples in the mini-batch, y_{ic} is an indicator of whether sample i belongs to class c , $\hat{p}_{ic}$ is the probability that the model assigns sample i to class c , and w_c is the weight assigned to class c . Each class weight is computed to be inversely proportional to the class frequency in the training data, mitigating the tendency of majority classes to dominate the loss and thereby impede the learning of minority classes.

Baseline 2 and end-to-end fine-tuning use the same multimodal fusion scheme and classification-head architecture, and differ only in whether the backbone parameters are updated. Comparing the two configurations therefore allows us to distinguish

between learning a classifier over fixed pretrained representations and adapting the multimodal representations themselves to the target task.

3.4 Experimental setup

3.4.1 Implementation details

All models were implemented using Python 3.11 and PyTorch 2.0, and trained on an NVIDIA TITAN RTX GPU with 24 GB of VRAM. The pre-trained backbones were loaded through the Hugging Face `transformers` library. We used `openai/clip-vit-base-patch32` for CLIP and the Korean-specialized `Bingsu/clip-vit-large-patch14-ko` for KoCLIP.

All learning-based neural configurations, including Baseline 2 and end-to-end fine-tuning, were trained using a shared optimization protocol to minimize the influence of differences in hyperparameter settings. The models were trained for up to 20 epochs using the Adam optimizer with a learning rate of 5×10^{-5} and a weight decay of 1×10^{-4} . In addition, a StepLR learning-rate scheduler was applied, with a step size of 5 and a decay factor γ of 0.1. Because the label distribution was non-uniform, the neural classifiers were optimized with class-weighted cross-entropy loss.

The effective batch size was fixed at 32 for both configurations. Baseline 2 used a batch size of 32 directly, whereas end-to-end fine-tuning used a batch size of 8 with 4-step gradient accumulation to accommodate GPU memory constraints. Early stopping was performed using quadratic weighted kappa (QWK) on the inner validation set, with a patience of 3 epochs.

For the ablation study (Results), two classical classifiers, an SVM and XGBoost, were trained on the same frozen KoCLIP embeddings. Both classifiers were evaluated under the same post-level grouped five-fold protocol as the neural configurations, with hyperparameters fixed to commonly used values rather than tuned by search. The SVM used an RBF kernel ($C = 1.0$, `gamma=scale`) with balanced class weights and embeddings standardized within each training fold; XGBoost used a multi-class softmax objective (`max_depth=6`, learning rate 0.05, subsample and column-subsample 0.8, at most 500 trees) with balanced sample weights, its tree count selected by early stopping on the same inner validation split as the neural models. Full configurations are provided in the accompanying code repository.

3.4.2 Evaluation protocol

All model configurations were evaluated using five-fold StratifiedGroupKFold with `post_id` used as the grouping variable. By splitting the data at the post level, multiple image-hashtag pairs generated from the same Instagram post were prevented from appearing in both the training and test partitions, thereby avoiding post-level information leakage.

For the learning-based configurations, each outer training partition was further divided into an inner training set and an inner validation set. This split used a GroupShuffleSplit based on `post_id`, with an inner-validation ratio of 0.1. The inner validation set was used only for early stopping and model selection, whereas class weights were computed exclusively from the label distribution of the inner training

599 set. The outer test fold was not used for model or hyperparameter selection, and was
600 reserved solely for final evaluation.

601 For the similarity-based configurations, the thresholds were optimized using only
602 the outer training partition of each fold, and the test partition of that fold was excluded
603 from the threshold-selection process. All evaluation metrics were computed on each of
604 the five folds and are reported as mean and standard deviation.

605

606 3.4.3 Evaluation metrics

607

608 As the task is a five-level ordinal classification problem, we use evaluation metrics
609 that capture overall predictive performance, class balance, and the ordered structure
610 of the labels. Accuracy measures the proportion of correctly classified samples. Macro
611 F1 assigns equal weight to the F1 score of each class and therefore reflects minority-
612 class performance independently of class frequency. Weighted F1 computes the average
613 class-wise F1 score weighted by the observed label distribution.

614 To quantify ordinal prediction error, we use mean absolute error (MAE), defined as

615

$$616 \text{MAE} = \frac{1}{N} \sum_{i=1}^N |y_i - \hat{y}_i|. \quad (11)$$

618

619 MAE represents the average distance error between the true and predicted classes.
620 In addition, we use the quadratic weighted kappa (QWK) to assess chance-corrected
621 ordinal agreement, which imposes larger penalties on errors with a greater distance
622 between the true and predicted classes.

623 Adjacent accuracy is defined as the proportion of samples whose predicted class is
624 identical to, or within one adjacent step of, the true class:

625

$$626 \text{AdjAcc} = \frac{1}{N} \sum_{i=1}^N \mathbb{I}(|y_i - \hat{y}_i| \leq 1). \quad (12)$$

628

629 By reporting accuracy alongside macro-F1, we examine whether predictions biased
630 toward the majority classes yield high overall accuracy while degrading balanced per-
631 formance across classes. MAE, QWK, and adjacent accuracy additionally evaluate
632 the magnitude and structure of ordinal errors that are not revealed by exact-class
633 agreement alone.

635

636 3.5 Statistical analysis

637

638 To determine whether observed performance differences between models could be
639 attributed to chance, we tested a set of theoretically motivated model compar-
640 isons. All analyses were based on the out-of-fold predictions generated through 5-fold
641 `StratifiedGroupKFold`. Accordingly, every sample in the dataset was evaluated
642 exactly once by a model that had not been trained on that sample, while all image-
643 hashtag pairs originating from the same Instagram post were assigned to the same
644 fold.

For all evaluation metrics—accuracy, macro-F1, quadratic weighted kappa (QWK), and mean absolute error (MAE)—we estimated 95% confidence intervals for the between-model differences using a post-level cluster bootstrap with 2,000 resamples. Resampling was performed at the post level (`post_id`) rather than at the individual image–hashtag-pair level because multiple pairs derived from the same post are not statistically independent. Post-level resampling therefore accounts for the grouped structure of the dataset and follows the same rationale as the group-based cross-validation procedure used to prevent data leakage. A difference was treated as statistically distinguishable when its 95% confidence interval did not include zero. For accuracy, we additionally performed McNemar’s test, which directly compares the two models’ correctness on the same samples. The bootstrap confidence intervals and the McNemar p -values thus provide complementary evidence.

We restricted the tests to three key model comparisons that were defined prior to inspecting the analysis results, based on our research questions RQ1–RQ3: (i) the effect of a Korean-specific multimodal backbone (fine-tuned KoCLIP vs. fine-tuned CLIP), (ii) the effect of fine-tuning relative to a frozen backbone (fine-tuned KoCLIP vs. frozen-backbone KoCLIP), and (iii) end-to-end fine-tuning versus a frozen-feature XGBoost classifier, both starting from the same pre-trained KoCLIP backbone (fine-tuned KoCLIP vs. KoCLIP + XGBoost). To control the family-wise error rate, we applied Holm correction to the three McNemar p -values, while the bootstrap confidence intervals are reported as supplementary effect-size estimates for each metric. The similarity-based models serve as reference conditions that do not employ a trainable classifier; since our statistical analysis focuses on learning-based comparisons that isolate the effects of backbone choice, fine-tuning, and classification strategy, the similarity-based results are reported separately using descriptive statistics.

3.6 Ethics statement and use of AI tools

This study did not involve human research participants. It analysed images, hashtags and captions from posts that Instagram users had made publicly available, with no interaction with, or intervention on, the account holders, and no participant was recruited. Under Article 13 of the Enforcement Rule of the Bioethics and Safety Act of the Republic of Korea, research that uses only information disclosed to the general public, without direct manipulation of the individuals or their environment, falls within the category exempt from institutional review board review; no approval or exemption determination was obtained from an institutional review board for this study, and the Institutional Review Board of Sungshin Women’s University reviews prospective research protocols only. Informed consent from the account holders was not sought. Only the post image, its hashtags, and its caption were used; account identifiers were not collected or used, captions were stripped of user mentions and URLs, and the released dataset refers to posts by anonymised identifiers. Because the images may depict identifiable individuals, they are not published or deposited in any public repository and are shared only under the conditions described in the Data availability statement; no facial recognition or other identification of individuals was performed. The four undergraduate research assistants who performed the validation labeling did so as members of the research team, not as research participants. All procedures were

691 carried out in accordance with relevant guidelines and regulations. No identifiable
692 individuals appear in the figures, which use photographs taken by the authors.

693 OpenAI GPT-4o mini (`gpt-4o-mini`, accessed through the Structured Outputs
694 API) was used for the automated consistency labeling of the dataset, as described in
695 detail above. ChatGPT (OpenAI, GPT-4o) was used for English grammar correction
696 and language polishing of the manuscript. The authors reviewed and edited all AI-
697 generated outputs and take full responsibility for the content of the manuscript; no
698 AI tool was used to create or edit any figure.

699

700

701

4 Results

702

703

704

705

706

707

708

709

710

4.1 Reliability of GPT-4o mini annotations

711

712

713

714

715

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

4.2 Main results

732

733

734

735

736

Table 3 summarizes the stratified group five-fold cross-validation results of the six main model configurations, reported as mean and standard deviation. Among the six main configurations, the end-to-end fine-tuned KoCLIP model achieved the best performance across all major evaluation metrics. Fine-tuned KoCLIP obtained an accuracy of 0.458, a macro F1 score of 0.439, and a weighted F1 score of 0.457, while

Table 2 Reliability analysis of GPT-4o mini annotations

Comparison	Pairwise comparisons	Ordinal Krippendorff's α (95% CI)	Quadratic-weighted κ (range)	Exact agreement (%)	Adjacent agreement (%)	MAE
Human–Human	6	0.72 (0.69–0.75)	0.73 $\pm$ 0.03 (0.69–0.77)	52.4 $\pm$ 2.8	91.8 $\pm$ 1.5	0.57 $\pm$ 0.04
GPT-4o mini–Human	4	—	0.68 $\pm$ 0.03 (0.64–0.72)	49.6 $\pm$ 2.1	89.8 $\pm$ 1.6	0.63 $\pm$ 0.03

Table 3 Comparison of experimental results

Configuration	Backbone	Accuracy	Macro-F1	Weighted-F1	MAE	QWK	Adj-Acc
Baseline 1	CLIP	0.287	0.165	0.216	1.387	0.049	0.596
		± 0.007	± 0.004	± 0.005	± 0.009	± 0.016	± 0.009
	KoCLIP	0.291	0.201	0.252	1.133	0.227	0.674
Baseline 2	CLIP	± 0.006	± 0.004	± 0.005	± 0.017	± 0.021	± 0.005
		0.289	0.236	0.268	1.450	0.107	0.577
	KoCLIP	± 0.031	± 0.052	± 0.049	± 0.081	± 0.058	± 0.039
End-to-End	CLIP	0.322	0.293	0.317	1.348	0.155	0.628
		± 0.017	± 0.043	± 0.032	± 0.029	± 0.037	± 0.023
	KoCLIP	0.243	0.143	0.142	1.432	0.060	0.632
End-to-End	CLIP	± 0.018	± 0.055	± 0.052	± 0.144	± 0.041	± 0.038
		0.458	0.439	0.457	1.055	0.307	0.720
	KoCLIP	± 0.012	± 0.011	± 0.012	± 0.043	± 0.040	± 0.013

also outperforming the other configurations in MAE, QWK, and adjacent accuracy. In particular, the substantial improvement in macro F1 indicates that the model did not rely solely on the majority classes but produced relatively balanced predictions across the full set of consistency levels. Its lowest MAE and highest QWK further suggest that it captured not only exact class agreement but also the ordinal distance between the predicted and true classes more effectively.

Overall, configurations with a trainable classification module tended to outperform those that relied only on pretrained embedding similarity. Similarity-based prediction yielded low macro F1 scores for both backbones, indicating that pretrained image–text similarity alone is insufficient for distinguishing five levels of semantic consistency. Frozen-backbone classification improved both macro F1 and weighted F1 over similarity-based prediction, showing that a supervised nonlinear decision boundary can enhance class discrimination and class-balanced performance even when the multimodal representations remain fixed. However, these improvements were not consistently observed across all ordinal metrics, suggesting that introducing a classification head alone does not fully overcome the limitations of the pretrained representations.

The effect of end-to-end fine-tuning differed markedly across backbones. For KoCLIP, fine-tuning led to the largest performance gains across all metrics, whereas fine-tuning CLIP resulted in a performance decrease, with its macro F1 falling to 0.143. This contrast, observed under the same training protocol, suggests that the effectiveness of fine-tuning depends not only on whether the backbone parameters are updated but also on how well the pretrained backbone is aligned with the target language and task. In particular, the Korean-adapted KoCLIP backbone appears to

Table 4 Ablation study comparing end-to-end fine-tuning with classical classifiers on frozen KoCLIP embeddings. All values are reported as mean $\pm$ standard deviation across the five folds. For all listed metrics except MAE, higher is better; for MAE, lower is better. The best value in each column is shown in bold.

Model	Accuracy	Macro-F1	Weighted-F1	MAE	QWK
KoCLIP + SVM	0.443 $\pm$ 0.015	0.419 $\pm$ 0.015	0.443 $\pm$ 0.014	1.041 $\pm$ 0.019	0.320 $\pm$ 0.011
KoCLIP + XGBoost	0.460 $\pm$ 0.011	0.420 $\pm$ 0.013	0.451 $\pm$ 0.010	1.027 $\pm$ 0.020	0.311 $\pm$ 0.020
KoCLIP (end-to-end)	0.458 $\pm$ 0.012	0.439 $\pm$ 0.011	0.457 $\pm$ 0.012	1.055 $\pm$ 0.043	0.307 $\pm$ 0.040

provide a more suitable initial representation for learning graded semantic consistency between images and Korean hashtags.

These results provide answers to the three research questions posed in this study. First, the relatively low performance of similarity-based configurations shows that pretrained embedding similarity alone is insufficient to capture graded semantic consistency (RQ1). Second, introducing a trainable classification head improves performance over similarity-based prediction, confirming the need for a supervised decision boundary (RQ2). Third, end-to-end fine-tuning of a suitable backbone provides the largest additional performance gain, with this effect being particularly pronounced for KoCLIP (RQ3).

4.3 Ablation study: fine-tuning versus classical classifiers

In the main results, the end-to-end fine-tuned KoCLIP achieved the highest macro F1 among the six main configurations. However, additional analysis is needed to determine whether this improvement results from adapting the multimodal representations through backbone fine-tuning or whether comparable performance can be achieved simply by applying a stronger classifier to frozen KoCLIP embeddings. To examine this question, we trained two classical classifiers, SVM and XGBoost, on the same frozen KoCLIP embeddings, with the configurations given in Methods (Implementation details), and compared them with the end-to-end fine-tuned KoCLIP.

As shown in Table 4, the three models achieved relatively similar performance in terms of overall accuracy and QWK. In particular, XGBoost obtained nearly the same accuracy as fine-tuned KoCLIP, indicating that strong overall classification performance can be achieved by applying a nonlinear classifier to frozen KoCLIP representations. In contrast, fine-tuned KoCLIP outperformed both classical classifiers in macro F1, which gives equal weight to the performance of each class. This result suggests that, despite their similar aggregate accuracy, the models differ in how effectively they handle the five consistency levels.

To examine this difference in greater detail, Figure 2 compares the confusion matrices and class-wise precision, recall, and F1 scores of the three models. Fine-tuned KoCLIP showed the lowest variation in class-wise recall and the highest minimum recall, indicating the most balanced performance across the full set of classes. By contrast, XGBoost achieved high recall for Class 5, the most frequent class, but substantially lower recall for Class 1, which contains fewer samples. Given that the dataset is relatively concentrated in Classes 4 and 5 (Supplementary Table S2), the strong

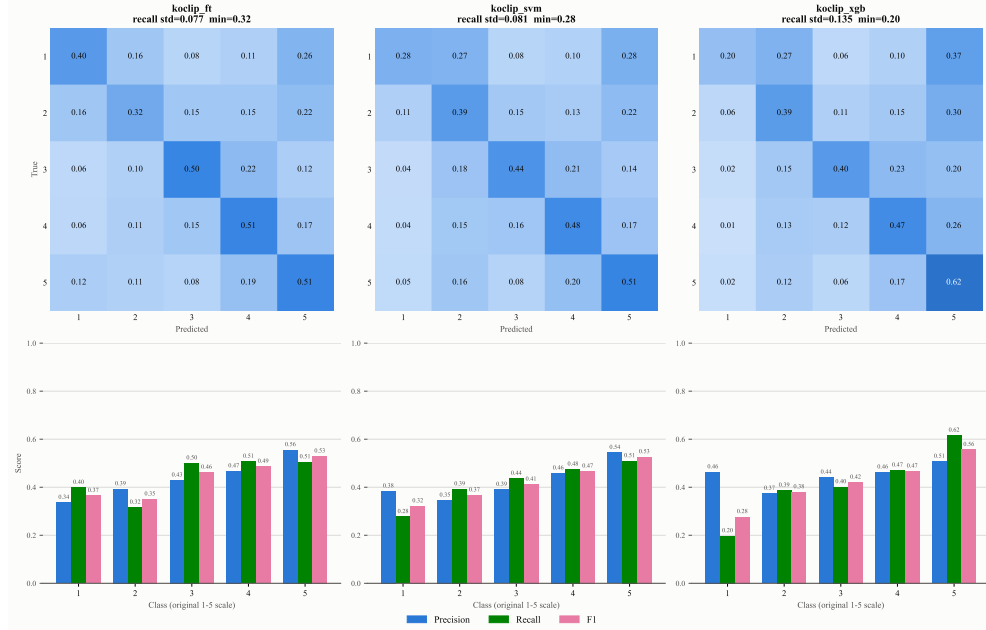


Fig. 2 Confusion matrices (top) and class-wise precision, recall, and F1 scores (bottom) for end-to-end fine-tuned KoCLIP, KoCLIP-SVM, and KoCLIP-XGBoost. Confusion-matrix values are row-normalized and therefore represent per-class recall.

overall accuracy of XGBoost appears to be influenced more heavily by its performance on the majority classes. This distinction has practical implications. Downstream uses of consistency prediction, such as filtering unrelated or exposure-driven hashtags in retrieval and recommendation, depend precisely on identifying the low-consistency classes; for such purposes, stable recall on these underrepresented classes can be as important as aggregate accuracy.

SVM exhibited less variation in class-wise recall than XGBoost and also achieved competitive QWK performance. However, it remained below fine-tuned KoCLIP in macro F1 and minimum recall. These results indicate that frozen KoCLIP embeddings already contain substantial task-relevant information for distinguishing semantic consistency between images and Korean hashtags, and that classical classifiers can effectively exploit this information. Because neither classifier was tuned for the task, their competitive aggregate performance primarily reflects the quality of the frozen KoCLIP representations rather than classifier engineering.

Overall, end-to-end fine-tuning did not consistently outperform the classical classifiers across every aggregate metric. Nevertheless, it provided the most balanced performance in terms of macro F1, weighted F1, minimum recall, and the stability of class-wise recall. This suggests that the primary contribution of fine-tuning lies not simply in increasing overall accuracy, but in jointly adapting the multimodal representations and the decision boundary to support more stable predictions across the imbalanced five-level consistency classes.

Table 5 Statistical significance testing for the key model pairs. For each metric we report the difference relative to the first model ($A - B$) and its 95% confidence interval; for accuracy, the Holm-adjusted McNemar p -value is also reported. For accuracy, macro-F1, and QWK a positive difference indicates that model A is superior, whereas for MAE a negative difference indicates that model A is superior. Confidence intervals were estimated with a post-level cluster bootstrap (2,000 resamples).

Comparison	Metric	Diff. ($A - B$)	95% CI	Holm-adj. p
KoCLIP (FT) vs. CLIP (FT)	Macro-F1	+0.285	[0.272, 0.298]	—
	QWK	+0.240	[0.204, 0.274]	—
	MAE	−0.378	[−0.416, −0.338]	—
	Accuracy	+0.214	[0.201, 0.228]	< .001
KoCLIP (FT) vs. KoCLIP (frozen)	Macro-F1	+0.133	[0.120, 0.147]	—
	QWK	+0.151	[0.119, 0.182]	—
	MAE	−0.293	[−0.333, −0.254]	—
	Accuracy	+0.135	[0.122, 0.149]	< .001
KoCLIP (FT) vs. KoCLIP + XGBoost	Macro-F1	+0.019	[0.008, 0.029]	—
	QWK	−0.004	[−0.029, 0.022]	—
	MAE	+0.0273	[0.0002, 0.0541]	—
	Accuracy	−0.002	[−0.012, 0.007]	.599

4.4 Statistical significance

Statistical testing followed the procedure described in Methods (Statistical analysis). The three model comparisons—fine-tuned KoCLIP versus fine-tuned CLIP, versus frozen-backbone KoCLIP, and versus KoCLIP + XGBoost—were fixed before the results were inspected.

Table 5 reports, for each of the three pre-specified comparisons, the metric differences with their bootstrap confidence intervals and Holm-adjusted McNemar p -values. The effect of the Korean-specific backbone was significant across all metrics: compared with fine-tuned CLIP, fine-tuned KoCLIP achieved a macro-F1 higher by 0.285 (CI [0.272, 0.298]) and an accuracy higher by 21.4 percentage points (Holm-adjusted $p < .001$), confirming that a Korean-specific multimodal backbone is essential for this task.

Fine-tuning was likewise significant across all metrics relative to the frozen backbone: fine-tuned KoCLIP exceeded frozen-backbone KoCLIP by 0.133 in macro-F1 and 13.5 percentage points in accuracy (Holm-adjusted $p < .001$), showing that adapting the entire backbone is consistently superior to training only the classification head.

The comparison with KoCLIP + XGBoost showed a different pattern. Accuracy and QWK did not differ significantly (accuracy McNemar Holm-adjusted $p = .599$; the QWK confidence interval included zero), whereas fine-tuned KoCLIP was significantly better in macro-F1 (difference +0.019, CI [0.008, 0.029]). The MAE difference of +0.0273 (CI [0.0002, 0.0541]) narrowly excluded zero but is very small and is better read as marginal. The two models are thus effectively equivalent in overall accuracy, while fine-tuned KoCLIP is more balanced across classes, including minority classes,

consistent with its lower recall variation and higher worst-class recall in the confusion-matrix analysis. End-to-end fine-tuning therefore contributes not by raising overall accuracy further but by maintaining balanced performance across the harder and less frequent classes.

4.5 Error analysis

Quantitative metrics show in which classes and to what extent the model makes errors, but they do not reveal in which types of image–hashtag pairs, or for what reasons, these errors occur. Here we analyze the misclassifications of the fine-tuned KoCLIP in terms of the distance between the true and predicted classes and the domain in which the errors occur, and we examine their causes through representative cases.

In terms of error magnitude, 48.3% of all misclassifications were adjacent errors, in which the true and predicted classes differed by one level, while 20.6% were large-distance errors, in which the classes differed by three or more levels. This indicates that, while a substantial portion of the misclassifications occurred between adjacent consistency levels, errors in which the prediction deviated substantially from the true label also accounted for a non-negligible proportion. As the two error types may reflect different causes and characteristics, they warrant separate interpretation.

A domain-level analysis showed that large-distance errors were relatively frequent in the Entertainment, Travel/Place, and Tech/Product domains, whereas their proportion was comparatively low in the Fashion and Beauty/Shopping domains. This difference is likely related to how the meaning of a hashtag is visually grounded in the image within each domain. Fashion- and beauty-related hashtags often refer to physical objects that appear directly in the image, whereas hashtags in the Entertainment, Tech/Product, and Travel/Place domains more often refer to specific people, product models, or places that require fine-grained identification or external knowledge. Figure 3 illustrates this contrast. In (A), a concrete visual object (cherry blossoms) is clearly present in the image, and the model accurately predicts the consistency of the corresponding hashtags. In (B), by contrast, the post shows the retail packaging of a specific smartphone model; although the hashtags directly name the depicted product (e.g., #iphone16, #iphone16promax), judging their consistency requires identifying the exact product model from packaging cues, and the model assigns the lowest consistency level to every product hashtag of the post (GT=5, Pred=1), so that the predictions deviate maximally from the true consistency. This shows that large-distance errors tend to occur when judging a hashtag requires fine-grained identification or external knowledge, rather than recognition of a generic object category.

This tendency is confirmed at the level of individual hashtags (Supplementary Note 2): visually concrete hashtags such as #메이크업 (makeup) and #단풍 (autumn foliage) had misclassification rates of only 5% and 13%, whereas the cultural slang #JMT (88%), the abstract seasonal mood #여름 (summer; 86%), and the drama title #참교육 (74%) were misclassified far more often, reflecting cultural-context, metaphorical, and external-knowledge gaps rather than random error.

In contrast, adjacent errors were not concentrated in particular domains but were distributed relatively evenly across domains. This suggests that adjacent errors arise

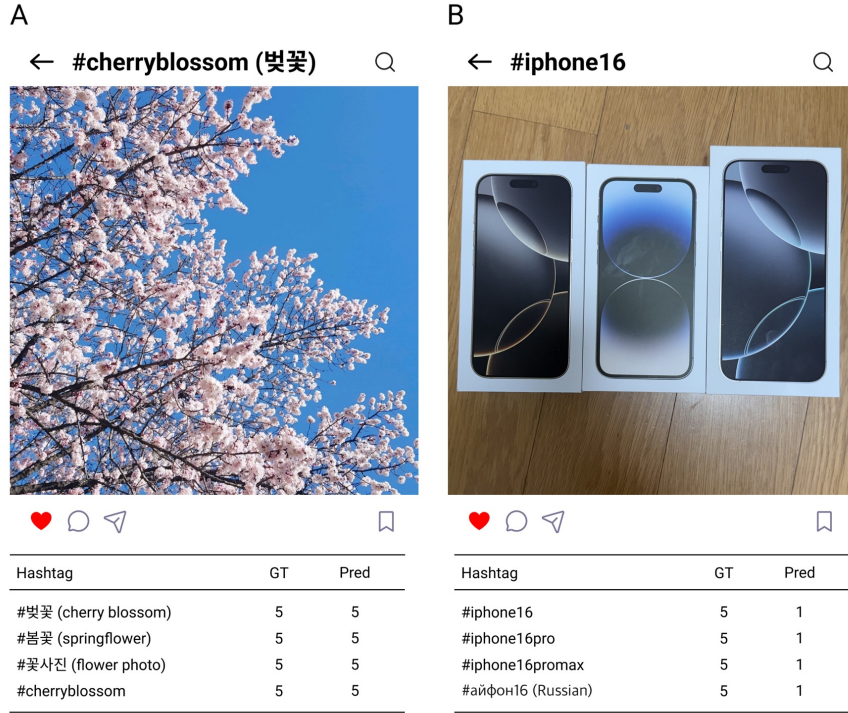


Fig. 3 Qualitative comparison of prediction behavior by visual groundability. Each panel corresponds to a representative Instagram post from our dataset and lists the model’s predictions (GT: ground-truth label; Pred: predicted label) for representative hashtags of that post. Because the original third-party photographs cannot be republished under the journal’s open license, both panels are illustrated with photographs taken by the authors that depict the same visual content as the original posts. (A) Season/Nature post: the concrete visual object (cherry blossoms) is directly present in the image, and the model accurately predicts high consistency for the object-level hashtags. (B) Tech/Product post, a photograph of iPhone retail packaging: although the hashtags directly name the depicted product, judging their consistency requires identifying the exact product model from packaging cues; the model fails to make this fine-grained identification and assigns the lowest consistency level to every product hashtag of the post (GT=5, Pred=1). The original post images belong to the dataset described in the Data availability statement. Photo credits: (A) Jangmin Oh; (B) Jangmin Oh.

not from the visual characteristics of a specific domain, but rather from the combined effect of the boundary ambiguity among the five consistency levels and the model’s ability to make fine-grained distinctions. In particular, for emotional or context-dependent hashtags, two adjacent labels may both be reasonable judgments, so such errors should be interpreted separately from large-distance errors.

Taken together, the failures of fine-tuned KoCLIP arise from several distinct factors rather than a single cause. Large-distance errors occurred mainly when the referent of a hashtag was not visually manifested in the image (specific people, products, or places) or required cultural context or external knowledge (slang, drama titles), whereas adjacent errors reflect the inherent boundary ambiguity among the five levels, where two

neighboring labels may both be reasonable for emotional or context-dependent hashtags. The accuracy of 45.8% is therefore better read not simply as a deficiency of the model but as reflecting the multifaceted difficulty of a task in which hashtags encode affective, cultural, and contextual meaning on which even humans do not fully agree.

5 Discussion

Our findings indicate that semantic consistency between images and hashtags on social media is not a simple binary property, but a graded and multi-dimensional construct involving literal object correspondence, affective association, contextual relevance, and culturally grounded interpretation. This observation is consistent with prior studies showing that hashtags do not merely refer to objects directly visible in an image, but may also express emotions, attitudes, social context, and user intentions [2, 4]. Consequently, the semantic relevance of a hashtag cannot always be determined solely from observable visual cues.

The low class-balanced performance of the similarity-based configurations with both backbones indicates that distances learned for general image-text alignment do not correspond to the graded consistency of affective and context-dependent hashtags, in line with broader observations that vision-language models pretrained on descriptive image-text pairs struggle with informal, non-literal language in social-media settings [6]. A supervised decision boundary over frozen representations recovers class discrimination and class balance, but its gains did not extend uniformly to the ordinal metrics, so a classification head alone does not resolve the ordinal structure of the task.

The effect of end-to-end fine-tuning differed markedly across the two backbones. Fine-tuning KoCLIP produced the most balanced performance, whereas fine-tuning CLIP under the same training protocol resulted in lower performance. This contrast suggests that the effectiveness of fine-tuning depends not only on updating the backbone parameters, but also on whether the pretrained backbone provides an initial representation that is suitable for the target language and task. The Korean-adapted KoCLIP backbone appears to offer a more appropriate starting point for learning the linguistic and cultural characteristics of Korean hashtags. By contrast, the degradation observed for CLIP may be related to unstable adaptation, overfitting on the task-specific dataset, or a mismatch between its pretrained representations and the target language. Determining the precise mechanism behind this degradation would require further analysis of training dynamics, parameter changes, and the structure of the learned embedding space.

The ablation study qualifies the contribution of fine-tuning. That untuned SVM and XGBoost classifiers on frozen KoCLIP embeddings matched fine-tuned KoCLIP on aggregate metrics, with no statistically distinguishable difference from XGBoost in accuracy or QWK, shows that the pretrained KoCLIP representations already encode much of the information needed for this task. The significant macro-F1 advantage of fine-tuned KoCLIP, together with its higher minimum class recall, indicates that the

benefit of fine-tuning is not a uniform gain in every aggregate metric but a joint adaptation of representations and decision boundary that yields more stable predictions across the imbalanced five-level label space, including the underrepresented Class 1.

Model performance should also be interpreted in conjunction with the annotation reliability analysis. In the human validation study, exact agreement between GPT-4o mini and the human annotators remained below 50%, and the human annotators themselves agreed exactly on only 52.4% of pairs, with about 8% of human–human comparisons differing by more than one level. Adjacent scores could represent similarly plausible judgments, particularly for hashtags involving affective, metaphorical, or culturally contextualized meanings. These findings indicate that the boundaries among the five consistency categories are not entirely objective and that annotation uncertainty should be considered when interpreting exact model accuracy.

Nevertheless, human agreement and model accuracy measure different quantities and should not be compared as if they were directly equivalent. Agreement results should therefore not be treated as a numerical upper bound on model performance, and annotation ambiguity does not explain all prediction errors. The adjacent accuracy of 0.720 achieved by fine-tuned KoCLIP indicates that many predictions were either correct or within one level of the reference label. However, large-distance errors between low- and high-consistency classes were also observed for some samples. This suggests that the difficulty of the task arises not only from subjective boundaries between neighboring categories, but also from model limitations in handling low-frequency slang, abstract affective expressions, fine-grained object or brand recognition, and contextual meanings that are not directly visible in the image.

Overall, the results demonstrate that modeling semantic consistency in Korean social-media content requires the joint consideration of language-appropriate multimodal representations, supervised decision boundaries, and the ordinal structure of prediction errors. KoCLIP-based fine-tuning provided the most balanced class-wise performance, while frozen KoCLIP representations combined with strong classifiers remained competitive in overall performance. Future model development should therefore consider not only improvements in exact accuracy, but also class-wise balance, ordinal agreement, computational efficiency, and the incorporation of additional contextual information.

5.1 Limitations

Although this study provides a systematic empirical evaluation of graded semantic consistency prediction between images and Korean hashtags, it has several limitations:

- **Limited Contextual Information:** The current model uses only the post image and an individual hashtag as input. However, the meaning of a hashtag may depend on additional contextual information, including the full caption, co-occurring hashtags, text embedded in the image, posting time, social trends, and the user’s characteristic style of expression. In particular, affective or metaphorical expressions and culturally grounded slang are often difficult to interpret from a single image–hashtag pair alone. Hashtags referring to specific products or places may also

- require fine-grained visual recognition or external knowledge. Context-aware models that integrate these additional signals represent a promising direction for future research. However, the use of user-level information must be accompanied by careful consideration of privacy protection and ethical data use.
- **Semantic Ambiguity:** Some hashtags are inherently abstract and subjective, making their consistency with an image difficult to evaluate objectively. For example, expressions such as #행복 (happiness) and #감성샷 (aesthetic shot) depend more on atmosphere, emotion, and personal interpretation than on concrete objects visible in the image. This inherent subjectivity complicates the interpretation of exact accuracy against a single reference label and may limit reliable discrimination between neighboring consistency categories. Future research may consider human-in-the-loop evaluation, context-aware representations, or probabilistic labels that preserve disagreement among annotators. More rigorous modeling of subjective semantic judgments may also benefit from additional empirical insights from psychology and cognitive science.
 - **Temporal and Platform Generalization:** The dataset used in this study consists primarily of public Korean Instagram posts collected during a specific period. Although post-level grouping prevented the same post from appearing in both the training and test partitions, the study did not conduct temporally separated evaluation. The meanings and usage patterns of hashtags may change rapidly across platforms, user populations, and social trends. Therefore, the present findings should not be assumed to generalize directly to newly emerging expressions, other social-media platforms, or different cultural contexts. Future studies should conduct out-of-domain evaluation using temporally separated test sets, external datasets from other platforms, and previously unseen hashtags.
 - **Model Complexity:** End-to-end fine-tuning of large multimodal models such as KoCLIP requires substantial computational resources, which may hinder deployment in resource-constrained environments. This limitation could be alleviated by adopting parameter-efficient fine-tuning methods, including LoRA, DoRA, or Uni-LoRA [19–21], as well as model compression or knowledge distillation techniques.
 - **Reliance on Automated Labeling:** Although the expanded validation study demonstrated meaningful ordinal agreement between GPT-4o mini and four human annotators, the automated labels should not be regarded as definitive ground truth. GPT-4o mini–human exact agreement remained below 50%, and approximately 10% of comparisons differed by more than one consistency level. In addition, the human validation was conducted by undergraduate research assistants from a single research environment, which may not capture the full diversity of interpretations among Korean social-media users. Future studies should involve annotators with more diverse demographic and cultural backgrounds and may employ probabilistic or distributional labels to preserve disagreement in subjective cases.

1151 6 Conclusion

1152
1153 User-generated hashtags on image-centric social media often diverge from literal
1154 visual content and instead express emotion, context, and culturally grounded mean-
1155 ing. Formulating image–Korean-hashtag consistency as a five-level ordinal task, we
1156 found a clear gap between descriptive image–text alignment and this affective, non-
1157 literal usage: embedding similarity alone was insufficient, supervised classification
1158 over frozen representations improved class-balanced performance, and end-to-end
1159 fine-tuning helped only when the pretrained backbone suited the target language—
1160 fine-tuned KoCLIP was significantly better than fine-tuned CLIP and frozen-backbone
1161 KoCLIP on every metric, whereas the identical protocol degraded CLIP. Against the
1162 strongest classical baseline, XGBoost, the significant advantage of fine-tuning was con-
1163 fined to class-balanced performance. Because human annotators themselves agreed
1164 exactly on only about half of the validated pairs, the remaining errors reflect both
1165 the limits of current multimodal representations and the intrinsic ambiguity of fine-
1166 grained consistency judgments. Future work should integrate contextual signals such as
1167 captions, co-occurring hashtags, and external cultural knowledge; compare parameter-
1168 efficient adaptation strategies that combine the efficiency of frozen KoCLIP classifiers
1169 with the class balance of fine-tuning; and adopt distributional labels and temporally
1170 separated, cross-platform evaluation to test robustness to evolving language use.

1171
1172 **Data availability.** All non-image data underlying the results reported in this arti-
1173 cle are openly available. The complete annotated dataset (12,011 image–hashtag pairs
1174 with post identifiers, hashtags, and five-level consistency labels), the cross-validation
1175 fold assignments, the per-sample out-of-fold predictions of all eight model configu-
1176 rations, the full annotation materials (prompt text, scoring rubric, boundary-case
1177 rules, and output schema), and all source code are available on GitHub (<https://github.com/askjiyun/insta-multimodal-consistency>) and permanently archived on
1178 Zenodo (<https://doi.org/10.5281/zenodo.21703134>). Author-created data are released
1179 under the CC BY 4.0 licence and code under the MIT licence. These materials allow
1180 every reported metric, table, and statistical test to be recomputed without access to
1181 the images.

1182
1183 The 1,673 original Instagram images are not publicly available because copyright
1184 in the photographs remains with the individual Instagram users who posted them,
1185 who have not licensed their redistribution, and because many of the images depict
1186 identifiable private individuals. The images are required only to retrain the models
1187 from scratch. The image archive is available from the corresponding author on reason-
1188 able request for non-commercial academic research; requests are normally answered
1189 within 14 days, and recipients are required to agree not to redistribute the images,
1190 not to attempt to identify the depicted individuals, and to delete the images upon
1191 completion of their study. The complete image archive is available to the editors and
1192 reviewers of this article for the purpose of evaluation.

1193
1194 **Author contributions.** Jiyeon Oh and Jangmin Oh conceived and designed the
1195 study. Jiyeon Oh performed the experiments, analysed the data, and prepared the
1196 figures and tables. Jiyeon Oh and Jangmin Oh wrote the code and carried out the

computational work. Both authors drafted the manuscript, revised it critically for important intellectual content, and approved the final version.

Funding. This work was supported by the Sungshin Women’s University Research Grant of 2025.

Competing interests. The authors declare no competing interests.

References

- [1] 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>.
- [2] Giannoulakis, S. & Tsapatsoulis, N. Evaluating the descriptive power of instagram hashtags. *Journal of Innovation in Digital Ecosystems* **3**, 114–129 (2016). URL <https://www.sciencedirect.com/science/article/pii/S2352664516300141>.
- [3] Giannoulakis, S. & Tsapatsoulis, N. Filtering instagram hashtags through crowd-tagging and the hits algorithm. *IEEE Transactions on Computational Social Systems* **6**, 592–603 (2019).
- [4] 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).
- [5] Radford, A. *et al.* Learning transferable visual models from natural language supervision. *arXiv:2103.00020* (2021).
- [6] 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>.
- [7] jaketae. Koclip: Korean clip (contrastive language-image pretraining). <https://github.com/jaketae/koclip> (2024). Accessed: 2025-11-01.
- [8] 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*.
- [9] 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>.
- [10] 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>.

1243 [11] Xia, H. *et al.* A multidimensional framework for analyzing image–text consistency
1244 in social media. *Applied Sciences* (2026). URL [https://api.semanticscholar.org/](https://api.semanticscholar.org/CorpusID:287680322)
1245 [CorpusID:287680322](https://api.semanticscholar.org/CorpusID:287680322).
1246

1247 [12] Tan, Z. *et al.* *Large language models for data annotation and synthesis: A survey.*
1248 in *Proceedings of the 2024 Conference on Empirical Methods in Natural Language*
1249 *Processing*, 930–957 (Association for Computational Linguistics, Miami, Florida,
1250 USA, 2024). URL <https://aclanthology.org/2024.emnlp-main.54/>.
1251

1252 [13] Pavlovic, M. & Poesio, M. *The effectiveness of LLMs as annotators: A compar-*
1253 *ative overview and empirical analysis of direct representation.* in *Proceedings*
1254 *of the 3rd Workshop on Perspectivist Approaches to NLP (NLPerspectives) @*
1255 *LREC-COLING 2024*, 100–110 (ELRA and ICCL, Torino, Italia, 2024). URL
1256 <https://aclanthology.org/2024.nlperspectives-1.11/>.
1257

1258 [14] Fu, X. & Liu, W. *How reliable is multilingual LLM-as-a-judge?* in *Findings*
1259 *of the Association for Computational Linguistics: EMNLP 2025*, 11040–11053
1260 (Association for Computational Linguistics, Suzhou, China, 2025). URL <https://aclanthology.org/2025.findings-emnlp.587/>.
1261

1262 [15] Zhang, X. *et al.* GPT-4V(ision) as a generalist evaluator for vision–language tasks
1263 (2023). URL <https://arxiv.org/abs/2311.01361>. [arXiv:2311.01361](https://arxiv.org/abs/2311.01361).
1264

1265 [16] Hayes, A. F. & Krippendorff, K. Answering the call for a standard reliability
1266 measure for coding data. *Communication Methods and Measures* **1**, 77–89 (2007).
1267

1268 [17] Krippendorff, K. *Content Analysis: An Introduction to Its Methodology* 4th edn
1269 (SAGE Publications, Los Angeles, 2019).
1270

1271 [18] Cohen, J. Weighted kappa: Nominal scale agreement with provision for scaled
1272 disagreement or partial credit. *Psychological Bulletin* **70**, 213–220 (1968).
1273

1274 [19] Hu, E. J. *et al.* Lora: Low-rank adaptation of large language models.
1275 *arXiv:2106.09685* (2021). URL <https://arxiv.org/abs/2106.09685>.
1276

1277 [20] Liu, S.-Y. *et al.* Dora: Weight-decomposed low-rank adaptation.
1278 *arXiv:2402.09353* (2024). URL <https://arxiv.org/abs/2402.09353>.
1279

1280 [21] Li, K. *et al.* Uni-lora: One vector is all you need. *arXiv:2506.00799* (2025). URL
1281 <https://arxiv.org/abs/2506.00799>.
1282
1283
1284
1285
1286
1287
1288

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [supplementary.pdf](#)