Surfacing Minor Nuances in Subjective Expressions: Initial Evidence for Improved Output Quality in Collaborative Caption Revision

Takeru Isaka, Jiro Nagao, Iwaki Toshima
Human Informatics Labs., NTT, Inc.
{takeru.isaka, jiro.nagao, iwaki.toshima}@ntt.com

Abstract

Creative collaboration benefits when collaborators can absorb the nuances embedded in others' subjective expressions. We focus on especially difficult-to-notice nuances, which we call **minor nuances**, and provide initial empirical evidence that supporting their communication can improve creative collaboration outcomes. We then propose a commonsense-generation-based method that identifies minor nuances in subjective expressions and presents them as cues for collaborators. We evaluated the proposed method in a collaborative image-caption revision experiment designed to reflect brainstorming. Compared with a baseline in which collaborators freely explained the nuances embedded in their own subjective expressions, our method significantly improved output-quality ratings from human evaluators. These results provide initial evidence suggesting that explicitly sharing minor nuances supports richer creative expression by helping convey fine-grained aspects of senders' inner states that are difficult to communicate through ordinary explanations.

1 Introduction

In creative collaboration, collaborators incorporate others' ideas to refine their own ideas and produce better outcomes. Because such ideas are difficult to define objectively, they are often conveyed through subjective expressions. We refer to a collaborator who produces a subjective expression as the sender and a collaborator who interprets it as the receiver. Subjective expressions can contain nuances that the receiver can readily interpret and others that they find difficult to interpret. We refer to the former as **major nuances** and the latter as **minor nuances**. In this work, we provide initial empirical evidence that supporting the communication of minor nuances can improve creative collaboration outcomes. In what follows, we focus on the sender–receiver pair as the most basic unit of interaction in creative collaboration, considering only one-way communication from the sender to the receiver. Although other collaborators may influence such interactions in multiparty settings (Sawyer, 2003), these influences are beyond the scope of this study.

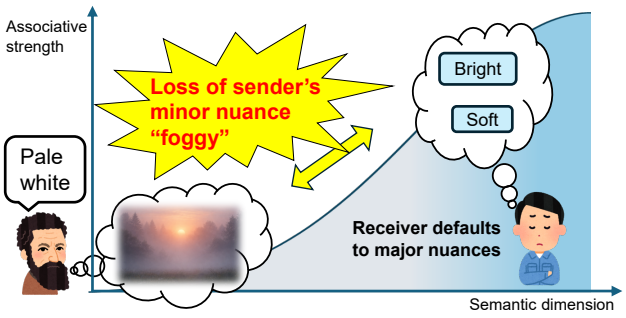


Figure 1: Motivation. Sender-specific, less-dominant (minor) nuances embedded in subjective expressions can be lost in creative collaboration when receivers converge on commonly inferred (major) nuances.

In creative collaboration, a sender weighs relatively minor nuances more heavily than in casual conversation (Sato, 1992, 1996). However, because major nuances are, by definition, more readily associated with a subjective expression, receivers tend to focus on them while overlooking minor nuances. If collaborators incorporate only such commonly associated nuances, their ideas may converge on common or predictable outcomes. For example, from the subjective expression "pale white," receivers easily infer major nuances such as "bright" or "soft," whereas the sender may intend a minor nuance such as "foggy" (Figure 1). In this case, the novel and useful color texture the sender intended to convey may be lost, replaced by a more conventional color. This view is consistent with prior work suggesting that creative cognition is associated with accessing less dominant or more remote associations (Mednick, 1962).

Qualitative studies of expert collaboration have also suggested that such collaboration can improve

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 5430–5446
November 6-10, 2026 ©2026 Association for Computational Linguistics

outcomes by eliciting minor nuances in others' subjective expressions (Howard, 1982; Ikuta, 2008; Isaka et al., 2024). However, the effect of explicitly communicating minor nuances in creative collaboration remains insufficiently examined. To address this gap, we formalize minor nuances with associative strength, as detailed in Section 3.

Our research question is therefore as follows:

Research question Does communicating minor nuances in subjective expressions improve the quality of outcomes in creative collaboration?

To answer this, we propose a communication framework for creative collaboration settings involving subjective verbal expressions, such as workplace discussions and brainstorming, in Section 3. The framework identifies nuances that are generally weakly associated with the sender's subjective expression but are relatively salient for the sender in creative collaboration, and it presents them to the receiver as cues to incorporate otherwise unnoticed minor nuances into their own ideas.

2 Related Work

2.1 Support for Communicating Subjective Expression

Many methods have been proposed to support communication of subjective expressions, often in tasks where language can be grounded in objective referents such as lexical definitions (Maekawa and Imai, 2023), physical objects (Narayan-Chen et al., 2019; Ichikawa and Higashinaka, 2022), diagrams (Mitsuda et al., 2022), and programming code (Sarkar et al., 2025). In these communication settings, the nuance of a subjective expression can be determined by identifying the correct referent or selecting standardized labels. Creative collaboration differs because the nuances in subjective expressions are difficult to ground in objective references.

Prior work has examined the functions and roles of subjective expressions in creative collaboration (Howard, 1982; Carter, 2004; Ikuta, 2008; Sato, 1996).

Clinical interviewing methods have explored how to help people verbalize private images and embodied experiences as subjective expressions, using this verbalization to support psychological care (Gendlin, 1997; Lawley and Tompkins, 2000; Petitmengin, 2006). However, to the best of our knowledge, no prior work has directly examined how sharing minor nuances affects creative outcomes.

2.2 Support for Enhancing Creativity

2.2.1 Idea Expansion through Relevant External Information

Most creativity-support methods help users explore ideas by providing external stimuli during individual or collaborative ideation. For individual ideation, such stimuli include images (Nakakoji et al., 1999; Wang et al., 2010; Petridis et al., 2021), word recommendations (Wise and Kenett, 2024), concrete examples (Siangliulue et al., 2016), and alternative expressions or rewrites of user input (Padmakumar and He, 2022; Chakrabarty et al., 2024). For collaborative ideation, prior methods present relationships among words or ideas expressed by collaborators (Sumi et al., 1996; Hori, 1994; Sumi et al., 1997) or conflicting viewpoints (Finke et al., 1992) as cues for reflection and recombination. Recent large-language model (LLM)-based methods further extend external-stimulus-based creativity support by generating fluent long-form text (Di Fede et al., 2022) and intricate visual content (Chakrabarty et al., 2024). In contrast to prior creativity-support studies, our work targets a different problem in creative collaboration: not how to add stimuli to expand ideas, but how to support the interpretive phase in which a receiver tries to understand the minor nuance in another collaborator's subjective expression.

2.2.2 Breaking Fixation through Random or Unrelated Stimuli

Creativity support methods based on lateral-thinking techniques deliberately introduce unrelated stimuli to help people break out of stalled problem frames (de Bono, 1970; Eno and Schmidt, 1975). Although this approach is valuable for overcoming ideation blocks, it focuses on a different phase of creative work from that targeted in this paper. In the interpretive phase of subjective expressions that we focus on, unrelated stimuli can become noise or distractions when the receiver tries to decipher the sender's intended nuance.

Importantly, these creativity-support methods and our method are not mutually exclusive; rather, they are complementary and can support different phases of a creative collaboration.

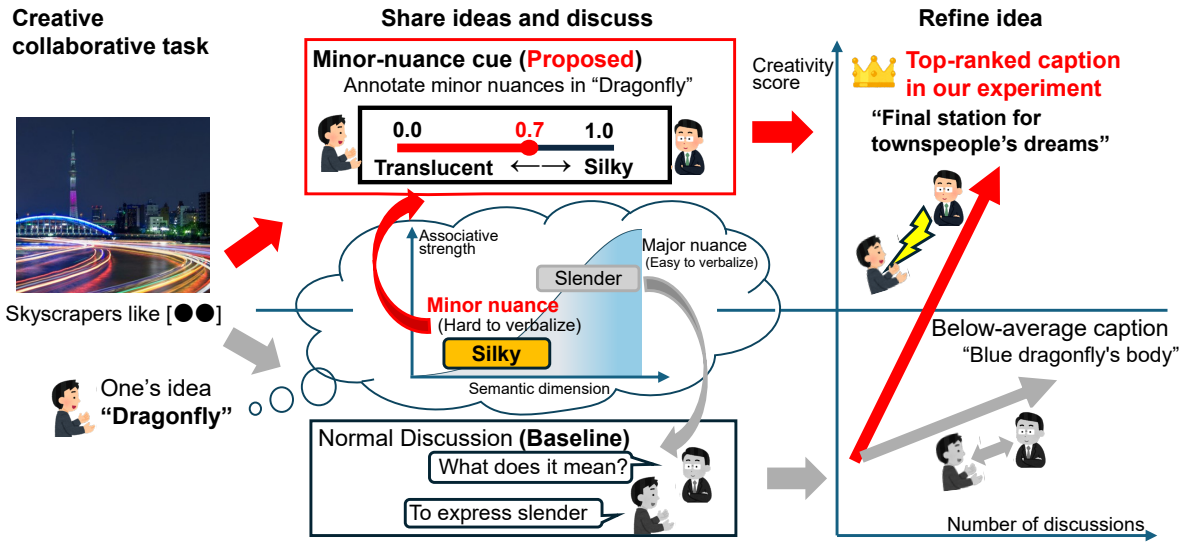


Figure 2: Overview of the proposed communication framework using actual examples collected in our collaborative image-captioning experiment. (A) In creative collaboration, a sender conveys their idea through a subjective expression. (B) The proposed method generates a pair of nuances (e.g., "translucent" and "silky") and shares the sender's annotation for the receiver to catch the sender's minor nuances. (C) This pair of minor nuances encourages the receiver to incorporate the sender's otherwise unnoticed nuance into their own idea.

3 Approach

3.1 Formulating Concept of Minor Nuances in Subjective Expressions

Here, we describe our concept to support receivers in noticing minor nuances in senders' subjective expressions. Following the view that concepts combine multiple components (Finke et al., 1992), we model a subjective expression X (e.g., "Pale white") as a weighted combination of candidate nuances:

$$X \approx \alpha_1 P_1 + \alpha_2 P_2 + \cdots + \alpha_n P_n,$$

where each P_i is a candidate nuance, and the weight α_i indicates how strongly the sender includes that nuance.

Major nuances are those that many people assign high weight to, whereas minor nuances are those that only particular senders assign high weight to. This explains why subjective expressions in creative collaboration are difficult to convey: the sender and receiver either do not share the same nuances underlying the expression or, even if they do, weigh those nuances differently.

Although the sender's latent minor nuance is difficult to directly recover, we aim to guide the receiver toward an associative region where the minor nuance becomes noticeable. This concept can be understood as directly enhancing creativity, defined as the ability to reach less dominant associations in Mednick's associative theory (Mednick, 1962).

3.2 Framework for Conveying Minor Nuances

Based on the concept described in Section 3.1, we propose a communication framework for conveying minor nuances in creative collaboration (Figure 2). The framework introduces a system mediator that helps bridge the gap between the sender's subjective expression and the receiver's interpretation.

After the sender produces a subjective expression, the mediator generates candidate nuances associated with that expression by using a commonsense generation model. It then operationalizes their associative strengths and determines two low-association nuances spanning different interpretive directions, as detailed in Section 4. The sender annotates which nuance better matches their intent, and the resulting annotation is shared with the receiver.

The receiver uses this annotation as a cue to reinterpret the sender's subjective expression and incorporate the sender's otherwise unnoticed nuance into their own creative output.

Our Hypothesis The proposed framework improves creative collaboration by conveying nuances that are generally weakly associated with a subjective expression but relatively salient for the sender.

Compared with conventional discussion, this framework helps the receiver incorporate the sender's otherwise overlooked nuances into their final output, thereby enhancing creativity.

Our contributions are as follows.

Contributions

- To the best of our knowledge, this study provides the first proof of concept explicitly testing whether communicating minor nuances in subjective expressions can improve creative collaboration outcomes.
- To conduct this test, we propose a novel method that helps collaborators catch the sender's minor nuances in subjective expressions, particularly those weakly associated with the expression yet relatively salient to the sender.
- We evaluate the proposed method in a collaborative image-captioning task conducted in Japanese and find that it significantly improves the quality of outcomes over a conventional communication baseline.

4 Method

Following Section 3, we implement a method that generates low-association nuance pairs for subjective expressions and asks the sender to annotate which of the pair better matches their intent. Figure 3 provides an overview of the method.

As a characteristic of human judgment, evaluating three or more ambiguous, subjective candidates makes people susceptible to response biases, such as central tendency and satisficing (Krosnick, 1991; Tourangeau et al., 2000), which can obscure within-person differences in relative importance. We therefore use a pairwise relative judgment between nuances: as shown in Figure 3(c), the sender is asked to choose which of the two words is closer to their intent. This is a standard format for eliciting personal characteristics while reducing the burden of multi-item subjective evaluation (Thurstone, 1927; Bradley and Terry, 1952).

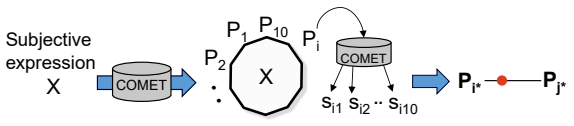


Figure 3: Overview of the low-association cue generation method. (a) Generate nuances P_1–P_{10}. (b) Reapply COMET to each P_i and generate secondary nuances S_{i1}–S_{i10}. (c) Select nuance pair (P_{i^*}, P_{j^*}) with fewest shared secondary nuances.

4.1 Generating Candidate Nuances

We first need a reasonable set of nuances associated with the input expression. We generate this set using COMET (Bosselut et al., 2019), which is an established baseline model for commonsense generation. Given an input language expression, COMET can produce a broad set of nuances that it can be plausibly associated with.

Later models, such as COMET-ATOMIC 2020, DiffuCOMET, and NovaCOMET (Hwang et al., 2021; Gao et al., 2024; West et al., 2023), extend COMET by providing broader coverage and leveraging stronger pretrained sequence-to-sequence models. Because these models follow the same commonsense-generation paradigm, demonstrating our method's usefulness on the original COMET would suggest it also transfers to later COMET variants. We therefore select the original COMET, which provides a stable proof-of-concept implementation for testing our research question.

For each input expression, we generate 10 nuances. In a preliminary inspection, requesting more than 10 outputs from COMET led to duplicate outputs rapidly increasing; see Appendix A.1 for details. We therefore used 10 outputs as a practical upper limit for obtaining unique nuance candidates from COMET.

4.2 Selecting a Low-Association Pair

The selected pair should satisfy two requirements:

(i) be minor nuances rather than central, readily inferred ones.

(ii) span different associative directions, reducing the chance of missing the sender's minor nuance.

Following theories that nuances with many shared features tend to be more central and readily recalled (Rosch and Mervis, 1975; Tversky, 1977;

Input	Nuance candidates	Selected pair
Dragonfly	Slender, Hard, Insect, Silky, Noisy, Flying, Translucent, Fast, Small, Useless	Silky, Translucent

Table 1: Example input and output of the proposed method. Appendix A.2 provides additional examples.

Giora, 1997), we define nuances that share fewer features with other nuances as minor nuances.

If we optimized only for requirement (i), we could select the two nuances that overlap the least with all other nuances on average. Instead, we choose the two nuances that mutually overlap the least to cover requirement (ii). This criterion approximates maximum coverage over the associative directions available to ordinary people, while still tending to avoid major nuances.

Our goal is to select two nuances whose evoked associations differ as much as possible. We therefore compare overlap between the association sets generated for each nuance using exact string matching within COMET's output space. This directly captures differences in associative content, rather than relying on word embeddings, which measure general distributional similarity in a corpus.

For each nuance P_i, we reapply COMET and generate a set of secondary nuances S_i as the set of associations evoked by P_i:

$$S_i = \{s_{i1}, s_{i2}, \ldots, s_{i10}\}. \quad (1)$$

If two nuances P_i and P_j evoke different associations, their corresponding sets S_i and S_j should overlap less.

We therefore compute the Jaccard distance between all pairs of S_i:

$$d_J(S_i, S_j) = 1 - \frac{|S_i \cap S_j|}{|S_i \cup S_j|}. \quad (2)$$

Then, we select the pair with the maximum distance:

$$(i^*, j^*) = \arg \max_{1 \le i < j \le 10} d_J(S_i, S_j), \quad (3)$$

and present (P_{i^*}, P_{j^*}) to the sender. Table 1 shows an example input-output of this method.

Because LLMs can also potentially generate minor nuances, we selected ChatGPT, the most widely used AI assistant (Barenholtz, 2025), as our LLM baseline and conducted two auxiliary comparisons, detailed in Appendix A.4. The results suggest that the proposed method can generate more minor nuances in subjective expressions than ChatGPT.

5 Experiment

5.1 Image Captioning Task

5.1.1 Task Design and Procedure

We conducted a participant experiment to test whether conveying minor nuances supports collaborative caption revision. We recruited 30 participants through a participant recruitment agency using non-probability sampling. Eligibility required no difficulty reading or writing Japanese. Participants were 20 women and 10 men aged 20–60 years (mean = 43). All participants were Japanese, and the experiment was conducted in Japanese. We therefore implemented a prototype system that could generate and present minor-nuance pairs in Japanese. [1]

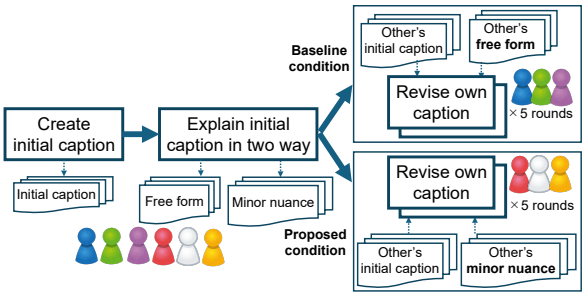


Figure 4: Overview of the experimental procedure.

We adopted collaborative image captioning as a creative collaboration task modeled on building on others' ideas during brainstorming. This task naturally elicits subjective expressions of visual impressions and produces relatively short textual outputs that human evaluators can readily compare.

Then, we selected the sender's free-form explanation of the caption as our baseline for two reasons. First, free-form explanations are likely to mainly convey the sender's major nuances. Second, in brainstorming—an intended use case for the proposed framework—senders often explain their subjective expressions in their own words to help receivers understand them. Thus, free-form sender explanations provide a valid baseline for ordinary communication in this experiment.

Figure 4 summarizes the experimental procedure. We conducted five sessions of a six-round image-captioning task, with six participants per session. We assigned the 30 participants to either the baseline or proposed condition (15 per condition), and each session included 3 participants from each condition, as illustrated in Table 2.

Each session used one image and a partially written caption with a blank (e.g., *Skyscrapers like [..]*). We manually selected the images from an image-caption dataset (Chen et al., 2015) and a photographer's portfolio (Kimoto, n.d.) so that each image

[1] Implementation details of the prototype system and screenshots of the experiment interface are provided in Appendix A.3 and A.6, respectively.

Session	Baseline	Proposed
1	1 (H), 3 (M), 5 (L)	2 (M), 4 (H), 6 (L)
2	7 (H), 8 (M), 9 (L)	10 (M), 11 (M), 12 (M)
3	13 (H), 15 (H), 16 (L)	14 (M), 17 (L), 18 (L)
4	19 (H), 20 (M), 24 (M)	21 (H), 22 (L), 23 (H)
5	25 (H), 26 (L), 27 (M)	28 (M), 29 (M), 30 (L)

Table 2: Participant assignments by session and condition. Each entry shows participant ID (RAT cluster). L, M, and H denote the low-, middle-, and high-score clusters defined from the distribution of all participants' Remote Associates Test (RAT) scores (Section 5.1.2).

contained multiple salient entities and supported more than one plausible interpretation.

In each session, participants first wrote an initial caption and then revised it five times, referring to others' initial captions and feedback. Thus, each participant submitted six captions in total: one initial and five revised. Each round of caption writing or revision had a 3-minute time limit, based on a pilot test in which five participants who did not take part in the main experiment completed the task comfortably and without fatigue. No participant in the main experiment was required to stop writing because of the time limit. Prior to the main session, participants completed a practice session to familiarize themselves with the task procedure.

5.1.2 Remote Associates Test

Before the captioning task, we administered the Japanese Remote Associates Test (RAT) (Orita et al., 2018). The RAT measures each participant's ability to find connections among seemingly weakly related concepts, a fundamental component of creative thinking (Mednick, 1962). Because our method presents nuances weakly associated with a subjective expression, we used RAT scores to balance this ability across conditions and examine whether it influenced the method' s effects in Section 7.

5.1.3 Creation and Explanation of Initial Caption

We instructed participants to complete the caption by filling in the blank with a phrase that described the image as detailed and figuratively as possible, while keeping the caption to within five words. This constraint was intended to encourage participants to embed nuance in concise expressions and make the subsequent explanations useful for extracting that nuance.

After submitting the initial caption, all participants explained it in two ways, regardless of condition. First, they wrote a free-form explanation. Then, they rated which of the pair of minor nuances presented by the proposed method was closer to the intended nuance using a slider ranging from 0 to 1. Because COMET's limitations can lead to nonsensical outputs for multiword inputs, participants selected up to two keywords from their initial captions, and a minor-nuance pair was generated separately for each keyword.

5.1.4 Iterative Revision and Subjective Evaluation

Participants revised their captions over five rounds, referring in each round only to the initial caption and associated feedback of a different one of the other five participants in the same session. These initial captions were created before any condition-specific feedback was presented; other participants' revised captions were never shown. The feedback shown to each participant depended on their assigned condition: baseline participants saw the caption author's free-form explanation, whereas proposed-condition participants saw the author's minor-nuance annotations. We instructed participants that they could leave their previous caption unchanged if they thought it was already optimal.

After the five rounds, participants rated their satisfaction with their final caption and the usefulness of each round's feedback they had viewed, using 5-point Likert scales.

5.2 Ranking of Submitted Captions

Following the basic rationale of the consensual assessment technique (Amabile, 1982), we aggregated independent subjective judgments from multiple general readers to assess perceived caption quality. For each session, submitted captions were ranked by 6–8 evaluators drawn from the same 30-participant pool described in Section 5.1.1. Evaluators were instructed to assign higher rankings to captions that described the image in greater detail and more figuratively. They evaluated only captions from sessions in which they had not participated and whose images they had not seen beforehand. In those sessions, they were blinded to caption authorship and interactions among participants.

Before evaluation, we excluded 9 captions containing 15 words or more because they substantially exceeded the approximately 5-word limit. We assessed inter-rater agreement using Kendall's co-

efficient of concordance (W) for caption rankings separately for each image and revision round (mean W = .325).

6 Analysis Methods and Results of Experiment

6.1 Analysis Methods

Because the number of captions evaluated per session varied, we normalized each evaluator's ranking to a value between 0 and 100 using a Borda-style score (de Borda, 1781):

$$S(c) = 100 \cdot \frac{1}{R} \sum_{r=1}^{R} \frac{m - \mathrm{rank}_r(c)}{m-1}, \tag{4}$$

where c is a submitted caption, $\mathrm{rank}_r(c)$ is the rank assigned to caption c by evaluator r, m is the number of captions in the session, and R is the number of evaluators. A higher $S(c)$ indicates a higher creativity ranking.

Caption Quality Metrics We then calculated two creativity-improvement measures:

$$\Delta S_{\text{global}} = S(c_{\text{highest}}) - S(c_{\text{init}}), \tag{5}$$

$$\Delta S_{\text{step}}(t) = S(c_t) - S(c_{\text{init}}), \quad t = 1, \ldots, 5. \tag{6}$$

where c_{init} denotes the participant's initial caption, c_t denotes the caption submitted after revision round t, and c_{highest} denotes the highest-scoring caption among the participant's six submissions in the session.

ΔS_{global} is computed once for each participant and captures improvement from the initial caption to the highest-scoring caption among that participant's submissions. ΔS_{step} is computed at the revision-round level and captures improvement from the initial caption after each feedback exposure.

Statistical Analysis For the nonparametric between-condition comparisons, we used the Brunner–Munzel test and report Cliff's δ as the effect size. Positive Cliff's δ indicates that the proposed condition was higher than the baseline condition.

Since ΔS_{step} contains repeated observations from the same participant, we used generalized estimating equation (GEE) models (Liang and Zeger, 1986) with participant-level clustering for the primary analysis of this measure. We set the significance level at $\alpha = .05$.

Metric	Proposed	Baseline	p	δ
ΔS_{step}	**5.13** [0.10, 10.15]	−2.78 [-6.99, 1.43]	**.009**	.26
ΔS_{global}	**37.95** [27.13, 48.77]	23.07 [16.97, 29.17]	**.026**	.49
GEE for ΔS_{step}: $\beta_{\mathrm{P-B}} = 7.91, p = .192$				

Table 3: Comparison of creativity-score improvements between proposed and baseline conditions. Values in brackets indicate 95% confidence intervals for each condition mean. δ denotes Cliff's δ. Bottom row presents results of GEE analysis of ΔS_{step} with participant-level clustering. ; $\beta_{\mathrm{P-B}}$ represents the estimated condition effect in direction Proposed − Baseline. Higher means and statistically significant p values are shown in bold.

Metric	Proposed	Baseline	p	δ
Caption satisfaction	**4.00** [3.59, 4.41]	3.58 [3.21, 3.96]	.159	.298
Feedback usefulness	2.37 [2.02, 2.72]	**3.02** [2.66, 3.38]	**.020**	−.506

Table 4: Comparison of subjective ratings between the proposed and baseline conditions.

6.2 Results: Effects on Caption Quality

We analyzed data from 27 participants after excluding three participants: one who did not complete the task and two who showed signs of drowsiness or difficulty concentrating early in the experiment.

As shown in Table 3, the proposed condition had a significantly higher ΔS_{global} than the baseline condition. The mean ΔS_{step} was also higher in the proposed condition, but the GEE analysis accounting for within-participant dependence did not show a statistically significant difference.

6.3 Results: Participants' Subjective Evaluation

As shown in Table 4, final-caption satisfaction was higher in the proposed condition than in the baseline. In contrast, feedback usefulness was significantly higher in the baseline condition than in the proposed condition. In Section 7.4, we discuss the mismatch between feedback usefulness and the observed improvements in creativity scores (Table 3).

7 Discussion

7.1 Effects of Conveying Minor Nuances

The experimental results support our hypothesis that conveying minor nuances improves creative collaboration. As shown in Table 3, the proposed condition showed a statistically significant advantage over the baseline in ΔS_{global}. These results suggest that the proposed method can improve creative outcomes when a receiver iteratively revises

their ideas using input from different senders.

7.2 Relationship with RAT Scores

7.2.1 Analysis by RAT Cluster

In the experimental results in Section 6, the proposed method showed a significantly larger ΔS_{global} than the baseline. However, they did not significantly differ in ΔS_{step}. Because the proposed method is designed to support catching less dominant associations, participants' associative abilities, which can be assessed using the RAT, may influence its effects.

We therefore conducted analyses by dividing participants into three RAT clusters—low, middle, and high—using thresholds at approximately $\pm 1\sigma$ from the mean, on the basis of the RAT scores obtained in Section 5.1.2. For each measure, we applied the Holm correction across the three condition comparisons in the low-, middle-, and high-RAT clusters.

Table 5 shows the analysis results. The proposed condition significantly improved over the baseline for both ΔS_{global} and ΔS_{step} in the low-RAT cluster, and for ΔS_{step} in the high-RAT cluster.

RAT	Metric	Proposed	Baseline	p	δ	GEE β	GEE p
Low	ΔS_{step}	**13.63** [7.36, 19.89]	−12.02 [-20.00, -4.04]	**<.001**	.78	25.65	**<.001**
	ΔS_{global}	**51.48** [33.25, 69.71]	19.27 [8.91, 29.63]	**.038**	.80	–	–
Middle	ΔS_{step}	−5.81 [-12.75, 1.14]	**2.66** [-4.08, 9.41]	.117	−.25	−8.47	.313
	ΔS_{global}	26.87 [13.65, 40.08]	**31.94** [23.80, 40.08]	.338	−.36	–	–
High	ΔS_{step}	**22.14** [13.33, 30.95]	0.26 [-5.85, 6.37]	**<.001**	.71	21.89	.060
	ΔS_{global}	**42.90** [30.08, 55.72]	19.02 [9.09, 28.94]	.190	1.00	–	–

Table 5: Comparison of creativity-score improvements within each RAT cluster. p values are Holm-adjusted across the three RAT clusters. For ΔS_{step}, results of GEE analysis are also shown. β represent estimated condition effects (Proposed − Baseline).

7.2.2 Interpretation of RAT-cluster Results

The following interpretation may explain why the proposed method showed a numerically greater advantage over the baseline in the low- and high-RAT clusters than in the middle-RAT cluster. High-RAT receivers may have been actively seeking minor nuances to improve their captions. For them, the proposed method was effective because it provided exactly the minor nuance they needed but had not noticed on their own. By contrast, free-form explanations from lower-RAT collaborators may have contained nuances that high-RAT receivers could already infer, making them less useful for further improving the caption.

Low-RAT receivers may have had difficulty extracting useful nuances from free-form explanations, especially those written by higher-RAT collaborators. The proposed method compressed useful minor nuances into a simpler form, making them easier to incorporate.

Middle-RAT receivers may have faced a different situation. Because they could, with some effort, understand captions from collaborators with similar RAT scores, they may have formed plausible interpretations of nuance. In the proposed condition, these self-generated interpretations may have competed with the minor nuances presented by the proposed method.

Based on these considerations, the high- and low-RAT results support our hypothesis, suggesting that explicitly presenting minor nuances can improve creativity more than nuances that receivers can readily infer. The middle-RAT results further suggest that future work should adapt the degree of minor nuance to the receiver's associative ability.

7.3 Comparison with Nuances of Different Associative Strengths

To examine whether the experimental results in Section 6.2 were specific to minor nuances, we compare the effect of minor nuances with three alternatives: nuances strongly associated with the original expression, unrelated nuances, and randomly selected nuances with uncontrolled associative strength.

First, the baseline free-form explanations in Section 5 can be regarded as strongly associative nuances, because they reflected the senders' own attempts to explain their captions. A semantic-distance analysis provided support for this interpretation, showing that these explanations were semantically closer to the original captions than the proposed minor nuances (Appendix A.4.2). Consistent with our hypothesis, the experimental results showed that these highly associative nuances were less effective than the proposed minor nuances (Section 6.2).

Second, unrelated nuances are unlikely to support interpretation of the sender's subjective expression and may even interfere with it, as discussed in Section 2. To examine this possibility, we conducted a post hoc analysis of 12 rounds in which participants reported in the questionnaire that the minor nuance introduced by the proposed method was incomprehensible or unrelated to the corresponding caption. We treated these rounds as pseudo-unrelated-stimulus cases and compared

changes in human evaluators' output-quality ratings for these rounds with those for the remaining proposed-method rounds. The ratings decreased in these pseudo-unrelated-stimulus rounds but increased in the other proposed-method rounds, as detailed in Appendix A.5.

Third, randomly associative nuances would likely be less effective than the proposed method because they would mix useful minor nuance with nuances that are either too highly associated or unrelated.

Taken together, these considerations suggest that the observed benefit of the proposed method cannot be reduced to the effect of other associative strengths.

7.4 Mismatch Between Creativity Gains and Perceived Usefulness

The proposed method improved creativity outcomes, yet participants rated it as less useful than free-form explanations. One possible interpretation is that the minor nuances did not work as immediately understandable explanations. Instead, they may have stimulted participants' own unverbalized interpretations of the subjective expressions, which were not consciously recognized as useful. This is consistent with research on implicit memory, priming, and tacit knowledge, in which prior stimuli can affect later performance without explicit awareness (Schacter, 1987; Polanyi, 1966).

Alternatively, minor nuances may simply have been confusing, irrelevant, or unhelpful. The present evidence does not distinguish these possibilities.

8 Conclusion and Future Work

8.1 Conclusion

This study provides initial empirical evidence that our approach to communicating minor nuances can improve caption quality in a collaborative image-caption revision task conducted in Japanese. As a proof of concept for our approach, we developed a commonsense-generation-based method that presents two low-association nuances derived from subjective expressions, along with sender annotations.

Compared with the conventional communication baseline, the proposed method showed a statistically significant advantage in maximum improvement from the initial caption over five revision rounds (ΔS_{global}). The difference in improvement averaged across rounds (ΔS_{step}) was not statistically significant overall. Both measures showed statistically significant advantages in the low-RAT cluster, suggesting that the proposed method can assist participants who have difficulty identifying remote associations. This finding further supports our hypothesis that conveying minor nuances helps receivers incorporate otherwise overlooked nuances into their creative outputs.

Taken together, these findings provide initial empirical evidence suggesting that our approach can facilitate a finer-grained understanding of others' subjective expressions—a fundamental process shared across forms of language-based creative collaboration—and enhance the creativity of the resulting outputs.

8.2 Future Work

Validating Minor Nuances Developing an appropriate procedure for assessing whether generated nuances are relevant to senders' intended meanings yet difficult to associate with their original expressions remains an important methodological challenge. Possible approaches include senders' relevance ratings, free-association tasks conducted before the generated nuances are presented, and semi-structured interviews with senders.

Isolating Component Effects Clarifying the respective contributions of minor-nuance content, presentation format, and sender annotation is an important challenge for future work. Addressing this requires controlled comparisons that include major-nuance and random-pair conditions with matched presentation formats and annotation requirements, conditions without sender annotations, and length-controlled free-form explanations, alongside measurements of cognitive load and processing effort.

Extending Nuance Generation Future implementations could explore other ConceptNet relations and compare or combine different language models to broaden the range of output nuances. Implementations that operate directly in the target language would also allow evaluation without translation.

Evaluating Broader Settings Larger studies should examine the framework in other languages and creative tasks involving subjective verbal expressions, including multiparty collaboration where participants exchange and develop ideas together.

9 Limitations

Study Scope and Translation The study analyzed data from 27 of 30 recruited Japanese participants in an image-caption revision task. The small sample, particularly within RAT clusters, limits generalization to other populations and creative tasks. Japanese–English translation before and after COMET generation may also have altered or failed to preserve fine-grained nuances in the original Japanese captions.

Nuance Validation and Coverage The operationalization of minor nuances was validated only indirectly. We did not directly verify whether the selected nuances were both relevant to senders' intended meanings and difficult to associate with their original expressions. Moreover, using only the ConceptNet HASPROPERTY relation restricted the generated nuances to attribute-like descriptions.

Presentation and Effort Confounds The experiment compared the proposed method as a whole with free-form explanations. The presented feedback differed in content, format, length, and the inclusion of sender annotations. Reading-related cognitive load and processing effort were not strictly controlled; completion within the three-minute limit does not establish comparable cognitive demands. All participants produced both explanations and annotations. The contributions of reflection during annotation and of displaying the resulting annotations were not isolated.

Missing Control Conditions We included no separate major-nuance, random-pair, or annotation-free control conditions. The post hoc analysis of cues perceived as unrelated does not establish superiority over randomly presented stimuli. Consequently, the observed gains cannot be attributed specifically to minor-nuance content.

Interpreting Usefulness Ratings The lower usefulness ratings also permit multiple interpretations. Some nuances may have required effort to interpret, whereas others may simply have been unhelpful. The available evidence does not distinguish these possibilities or establish the mechanism underlying the observed improvements. Qualitative follow-up methods, such as semi-structured interviews with receivers, could help the reasons for these ratings.

One-Way Communication Finally, real-world creative collaboration involves participants continuously influencing one another through the exchange and interpretation of ideas (Sawyer, 2003). Participants in our experiment referred only to others' fixed initial captions and associated feedback, so the study did not capture reciprocal exchanges through which ideas evolve among multiple collaborators.

Acknowledgments

We thank the anonymous reviewers for their insightful suggestions and discussions.

Ethics Statement

All participants provided written informed consent after receiving an explanation of the research purpose, collected data, and intended use. Participation was voluntary, and participants were informed that they could withdraw at any time without penalty. All collected data were anonymized before analysis, and no personally identifiable information was retained. The study protocol was reviewed and approved by the ethics review board of the Digital Twin Computing Lab., Human Informatics Labs., NTT, Inc. (Approval No.: DTP-ERB-2025-01).

References

Teresa M. Amabile. 1982. Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43(5):997–1013.

Anthropic. 2026. Claude Fable 5 and Claude Mythos 5. Published 2026-06-09; accessed 2026-10-01.

Limor Barenholtz. 2025. Top AI tools by web and app usage (August 2025 data). Published 2025-09-14; August 2025 usage data; accessed 2026-10-01.

Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. 2019. COMET: Commonsense transformers for automatic knowledge graph construction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4762–4779, Florence, Italy. Association for Computational Linguistics.

Ralph Allan Bradley and Milton E. Terry. 1952. Rank analysis of incomplete block designs: I. the method of paired comparisons. *Biometrika*, 39(3/4):324–345.

Ronald Carter. 2004. *Language and creativity: The art of common talk*. Routledge, London. Paperback ISBN: 9780415234498.

Tuhin Chakrabarty, Vishakh Padmakumar, Faeze Brahman, and Smaranda Muresan. 2024. Creativity support in the age of large language models: An empirical study involving professional writers. In *Proceedings of the 16th Conference on Creativity & Cognition*, C&C '24, pages 132–155, New York, NY, USA. Association for Computing Machinery.

Jianfu Chen, Polina Kuznetsova, David Warren, and Yejin Choi. 2015. Déjà image-captions: A corpus of expressive descriptions in repetition. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 504–514, Denver, Colorado. Association for Computational Linguistics.

Edward de Bono. 1970. *Lateral thinking: Creativity step by step*. Harper & Row, New York.

Jean-Charles de Borda. 1781. Mémoire sur les élections au scrutin. *Histoire de l'Académie Royale des Sciences*. Annual volume for 1781; printed publication is dated 1784.

DeepL SE. n.d. DeepL API documentation. Accessed 2026-10-01.

Giulia Di Fede, Davide Rocchesso, Steven P. Dow, and Salvatore Andolina. 2022. The Idea Machine: LLM-based expansion, rewriting, combination, and suggestion of ideas. In *Proceedings of the 14th Conference on Creativity and Cognition*, pages 623–627. Association for Computing Machinery.

Brian Eno and Peter Schmidt. 1975. Oblique strategies: Over one hundred worthwhile dilemmas.

Ronald A. Finke, Thomas B. Ward, and Steven M. Smith. 1992. *Creative cognition: Theory, research, and applications*. The MIT Press.

Silin Gao, Mete Ismayilzada, Mengjie Zhao, Hiromi Wakaki, Yuki Mitsufuji, and Antoine Bosselut. 2024. DiffuCOMET: Contextual commonsense knowledge diffusion. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4809–4831, Bangkok, Thailand. Association for Computational Linguistics.

Eugene T. Gendlin. 1997. *Experiencing and the creation of meaning: A philosophical and psychological approach to the subjective*. Northwestern University Press.

Rachel Giora. 1997. Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics*, 8(3):183–206.

Koichi Hori. 1994. A system for aiding creative concept formation. *IEEE Transactions on Systems, Man, and Cybernetics*, 24(6):882–894.

Vernon Alfred Howard. 1982. *Artistry: The work of artists*. Hackett Publishing.

Jena D. Hwang, Chandra Bhagavatula, Ronan Le Bras, Jeff Da, Keisuke Sakaguchi, Antoine Bosselut, and Yejin Choi. 2021. (COMET-) ATOMIC 2020: On symbolic and neural commonsense knowledge graphs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 6384–6392.

Takuma Ichikawa and Ryuichiro Higashinaka. 2022. Analysis of dialogue in human-human collaboration in Minecraft. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4051–4059, Marseille, France. European Language Resources Association.

Kumiko Ikuta. 2008. The role of "craft language" in learning "Waza". In Satinder Gill, editor, *Cognition, Communication and Interaction: Transdisciplinary Perspectives on Interactive Technology*, pages 405–414. Springer, London.

Takeru Isaka, Atsushi Otsuka, and Iwaki Toshima. 2024. Analysis of sensation-transfer dialogues in motorsports. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 876–886, Torino, Italia. ELRA and ICCL.

Yuya Kimoto. n.d. Yuya's portfolio. Accessed 2026-01-06.

Jon A. Krosnick. 1991. Response strategies for coping with the cognitive demands of attitude measures in surveys. *Applied Cognitive Psychology*, 5(3):213–236.

Takio Kurita, Toshikazu Kato, Ikumi Fukuda, and Ayumi Sakakura. 1992. Sense retrieval on a image database of full color paintings. *IPSJ Journal*, 33(11):1373–1383. In Japanese.

James Lawley and Penny Tompkins. 2000. *Metaphors in mind: Transformation through symbolic modelling*. The Developing Company Press.

Kung-Yee Liang and Scott L. Zeger. 1986. Longitudinal data analysis using generalized linear models. *Biometrika*, 73(1):13–22.

Tomoyuki Maekawa and Michita Imai. 2023. Identifying statements crucial for awareness of interpretive nonsense to prevent communication breakdowns. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12550–12566, Singapore. Association for Computational Linguistics.

Sarnoff A. Mednick. 1962. The associative basis of the creative process. *Psychological Review*, 69(3):220–232.

Koh Mitsuda, Ryuichiro Higashinaka, Yuhei Oga, and Sen Yoshida. 2022. Dialogue collection for recording the process of building common ground in a collaborative task. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5749–5758, Marseille, France. European Language Resources Association.

Kumiyo Nakakoji, Yasuhiro Yamamoto, and Masao Ohira. 1999. A framework that supports collective creativity in design using visual images. In *Proceedings of the 3rd Conference on Creativity & Cognition*, C&C '99, pages 166–173, New York, NY, USA. Association for Computing Machinery.

Anjali Narayan-Chen, Prashant Jayannavar, and Julia Hockenmaier. 2019. Collaborative dialogue in Minecraft. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5405–5415, Florence, Italy. Association for Computational Linguistics.

OpenAI. n.d. GPT-5 mini model. Accessed 2026-10-01.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, and 261 others. 2024. GPT-4 technical report. *Preprint*, arXiv:2303.08774. Version 6, revised 2024-03-04; first submitted 2023-03-15.

Ryo Orita, Masasi Hattori, and Yuki Nishida. 2018. Development of a Japanese remote associates task as insight problems. *The Japanese Journal of Psychology*, 89(4):376–386.

Vishakh Padmakumar and He He. 2022. Machine-in-the-loop rewriting for creative image captioning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 573–586, Seattle, United States. Association for Computational Linguistics.

Claire Petitmengin. 2006. Describing one's subjective experience in the second person: An interview method for the science of consciousness. *Phenomenology and the Cognitive Sciences*, 5(3–4):229–269.

Savvas Petridis, Hijung Valentina Shin, and Lydia B. Chilton. 2021. SymbolFinder: Brainstorming diverse symbols using local semantic networks. In *The 34th Annual ACM Symposium on User Interface Software and Technology*, pages 385–399. Association for Computing Machinery.

Michael Polanyi. 1966. *The tacit dimension*. Doubleday, Garden City, NY.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992. Association for Computational Linguistics.

Eleanor Rosch and Carolyn B. Mervis. 1975. Family resemblances: Studies in the internal structure of categories. *Cognitive Psychology*, 7(4):573–605.

Rupak Sarkar, Neha Srikanth, Taylor Pellegrin, Rachel Rudinger, Claire Bonial, and Philip Resnik. 2025. Understanding common ground misalignment in goal-oriented dialog: A case-study with Ubuntu chat logs. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3200–3215, Vienna, Austria. Association for Computational Linguistics.

Nobuo Sato. 1992. *Retorikku kankaku*. Kodansha, Tokyo. In Japanese.

Nobuo Sato. 1996. *Retorikku no imiron: Imi no dansei*. Kodansha, Tokyo. In Japanese.

R. Keith Sawyer. 2003. *Group creativity: Music, theater, collaboration*. Lawrence Erlbaum Associates, Mahwah, NJ.

Daniel L. Schacter. 1987. Implicit memory: History and current status. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 13(3):501–518.

Pao Siangliulue, Joel Chan, Steven P. Dow, and Krzysztof Z. Gajos. 2016. IdeaHound: Improving large-scale collaborative ideation with crowd-powered real-time semantic modeling. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*, pages 609–624. Association for Computing Machinery.

Isao Sonobe. 2021. sentence-bert-base-ja-mean-tokens-v2. Japanese Sentence-BERT model; accessed 2026-10-01.

Robyn Speer, Joshua Chin, and Catherine Havasi. 2017. ConceptNet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 31.

Yasuyuki Sumi, Koichi Hori, and Setsuo Ohsuga. 1997. Computer-aided thinking by mapping text-objects into metric spaces. *Artificial Intelligence*, 91(1):71–84.

Yasuyuki Sumi, Kazushi Nishimoto, and Kenji Mase. 1996. Facilitating human communications in personalized information spaces. In *Proceedings of the AAAI Workshop on Internet-Based Information Systems*, pages 123–129.

L. L. Thurstone. 1927. A law of comparative judgment. *Psychological Review*, 34(4):273–286.

Roger Tourangeau, Lance J. Rips, and Kenneth Rasinski. 2000. *The psychology of survey response*. Cambridge University Press, Cambridge.

Amos Tversky. 1977. Features of similarity. *Psychological Review*, 84(4):327–352.

Hao-Chuan Wang, Dan Cosley, and Susan R. Fussell. 2010. Idea Expander: Supporting group brainstorming with conversationally triggered visual thinking stimuli. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, pages 103–106. Association for Computing Machinery.

Peter West, Ronan Le Bras, Taylor Sorensen, Bill Yuchen Lin, Liwei Jiang, Ximing Lu, Khyathi Chandu, Jack Hessel, Ashutosh Baheti, Chandra Bhagavatula, and Yejin Choi. 2023. NovaCOMET: Open commonsense foundation models with symbolic knowledge distillation. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1127–1149, Singapore. Association for Computational Linguistics.

Talia A. Wise and Yoed N. Kenett. 2024. Sparking creativity: Encouraging creative idea generation through automatically generated word recommendations. *Behavior Research Methods*, 56(7):7939–7962.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot Arena. In *Advances in Neural Information Processing Systems*, volume 36, pages 46595–46623.

A Appendix

A.1 Number of Nuances Generated by COMET

In principle, COMET can generate an arbitrary number of nuances. However, increasing the number of outputs often results in many duplicate strings. We therefore conducted a preliminary analysis to determine a practical number of nuances for the experiment in Section 5, aiming to cover the widest possible range of COMET-generated nuances while keeping the experiment feasible.

We selected 30 impression words commonly used to describe images or paintings (Kurita et al., 1992), such as *cool*, *refreshing*, *luxurious*, and *sophisticated*. Each word was used as the input to COMET, and the number of requested outputs was varied. We then counted the number of unique string outputs for each input word and averaged the results across all 30 words.

As shown in Figure 5, the number of unique strings increased almost linearly up to around 10 requested outputs but then increased much more slowly. Requesting more outputs continued to yield few additional unique strings, and many of the extra outputs were duplicates. On the basis of this result, we set the number of nuances to 10 in the nuance generation procedure described in Section 4.

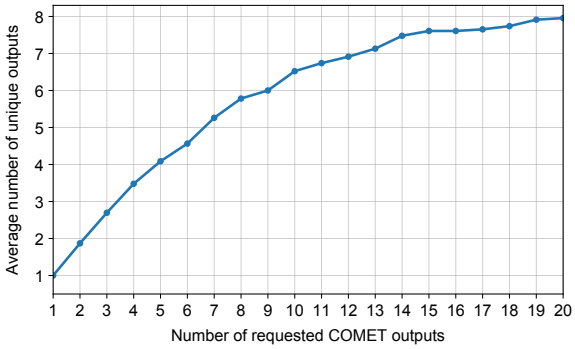


Figure 5: Preliminary analysis of the number of unique nuances generated by COMET. The horizontal and vertical axes indicate the number of outputs requested for COMET and the average number of unique outputs, respectively.

A.2 Additional Examples of Generated Nuances and Caption Revisions

A.2.1 Additional Examples of Generated Nuance Pairs

To illustrate the range of nuances generated by the proposed method beyond the examples in the main text, Table 6 shows additional examples of the input subjective expressions, COMET-generated nuance candidates, and the selected low-overlap nuance pairs. This method selects pairs in which each nuance is less readily associated with the input and the two evoke associations that are clearly different, as described in Section 4. Consequently, the selected nuance pairs tend to appear semantically incoherent from the input expression.

A.2.2 Illustrative Examples of Caption Revisions

Table 7 shows illustrative examples of caption revisions after participants referred to proposed nuances. For example, in one round, a receiver referred to another participant's initial caption, "Dragonfly," together with the proposed minor-nuance pair *Translucent* / *Silky*. The receiver then revised their own caption from "Magic creation" to "Shining magic," and the output-quality ratings improved by approximately three positions.

A.3 Prototype Implementation Details of Proposed Method

We describe the implementation details of the prototype system used in the experiment. The proposed method in Section 4 defines a general procedure for presenting low-association cues, but the experiment required a specific implementation capable

Input	Nuance candidates	Selected nuance pair
Magic items	Fun to touch, Cool, Fun, Beautiful, Dangerous, Unpredictable, Fun to borrow, Fun to use, Fun to watch, Fun to read	Beautiful / Unpredictable
Noble	Important, Right-wing, Good than poor, Arrogant, Right-handed, Moral, Selfish, Rich, Generous, Good	Arrogant / Right-handed
Lighthouse	Creepy, Beautiful, Romantic, Scary, Large, Dangerous, Dark, Impressive, Cool, Pretty	Large / Romantic

Table 6: Additional examples of input expressions, generated nuance candidates, and selected minor-nuance pairs.

Referenced caption	Referenced minor nuances	Caption revision	Approx. rank improvement
Dragonfly	Translucent / Silky	Magic creation → Shining magic	3 positions
Magic items	Beautiful / Unpredictable	Sword piercing night sky → Fairy's wand	3 positions
Noble	Arrogant / Right-handed	Elegant while looking back → King of the sky	6 positions
Lighthouse	Large / Romantic	Scarecrow king → Watchful woman	4 positions

Table 7: Illustrative examples of caption revisions after participants referred to minor nuances. Referenced caption indicates the other participant's initial caption that the participant referred to during revision. Referenced minor nuances indicate minor-nuance pairs generated by the proposed method referred to during revision. Rank improvement indicates the improvement in output-quality ratings after revision.

of processing Japanese subjective expressions and presenting cues to participants in Japanese.

Because COMET operates in English but the experiment was conducted in Japanese, we used machine translation before and after COMET generation. Specifically, the sender's Japanese subjective expression was first translated into English using the DeepL API (DeepL SE, n.d.). The translated expression was then input into COMET (Bosselut et al., 2019) using the HASPROPERTY relation, which specifies the relation between the input concept and the generated output. We selected HASPROPERTY because it represents a relation between two concepts, A and B, such that A has B as a property or can be described as B (Speer et al., 2017). We considered this the most direct relation for generating nuances that reflect how a subjective expression may be perceived.

The prototype then selected a low-overlap nuance pair on the basis of the cue-selection procedure described in Section 4.2. Finally, the selected nuance pair was translated back into Japanese using the DeepL API and presented to the participants.

In the prototype, we used the ConceptNet pretrained model from the original COMET implementation[2]. COMET provides multiple decoding options for generating candidate output phrases from an input concept and relation. For reproducibility, we used beam search rather than stochastic sampling, setting the beam size to 10 on the basis of the candidate-number analysis in Appendix A.1. Other parameters were left as the default settings of the original implementation.

A.4 Preliminary LLM Comparison and Semantic-Distance Analysis

We conducted two auxiliary analyses to examine whether the proposed method is more effective than LLMs at extracting minor nuances from subjective expressions.

A.4.1 Preliminary LLM Comparison

We compared COMET with ChatGPT-4o (OpenAI et al., 2024), a general-purpose LLM. For each of 30 impression words commonly used to describe images or paintings (Kurita et al., 1992), such as *cool*, *refreshing*, *luxurious*, and *sophisticated*, each model generated 10 candidate nuances that the word could convey.

For ChatGPT-4o, we used two generation prompts that differed in how strongly they encouraged minor nuances (Table 8). We then used ChatGPT-4o as an LLM judge (Zheng et al., 2023), assigning it the role of an expert in word associations to classify each candidate as easy or difficult to associate with the input word.

Because ChatGPT-4o both generated and evalu-

[2] https://github.com/atcbosselut/comet-commonsense. Accessed May 25, 2026.

ated its own candidates, this evaluation may involve circularity and model-specific bias. We therefore repeated the evaluation with Claude Fable 5 (Anthropic, 2026) using the same criteria. Under both judges, COMET generated the highest mean number of difficult-to-associate nuances (Table 9).

A.4.2 Semantic-Distance Analysis Using Experimental Data

We conducted a semantic-distance analysis using the experimental data. We embedded English translations of the submitted captions and three types of explanatory text—proposed-method annotations, free-form explanations, and LLM-generated nuances—using Sentence-BERT (Reimers and Gurevych, 2019) and computed their cosine distances from the corresponding captions. We generated the LLM nuances with GPT-5 mini (OpenAI, n.d.) using the prompt shown in Table 10, which instructed the model to approximate the proposed method's output by producing two nuances associated with each subjective expression but weakly related to each other.

We visualized the distance distributions using kernel density estimation (KDE). Greater distance indicates greater semantic separation from the original caption. As shown in Figure 6, the nuances generated by the proposed method had the rightmost peak, suggesting that they were less readily associated with the caption than the free-form explanations and LLM-generated nuances.

Because the experimental pipeline in Section 5 involved Japanese–English translation, we repeated the analysis in a Japanese embedding space to verify that the observed pattern was not specific to the English embedding space. Using a Japanese Sentence-BERT model (Sonobe, 2021), we obtained the same ordering of KDE peaks: free-form explanations (0.606), LLM-generated nuances (0.763), and proposed-method annotations (0.859).

These auxiliary analyses suggest that the proposed method better captures minor nuances in subjective expression than human free-form explanations or LLMs.

A.5 Comparison with Unrelated Cue

As a post hoc analysis, we examined rounds in the proposed condition in which participants' free-text comments suggested that the cue was unrelated to or difficult to connect with the original caption. We treated these rounds as perceived-unrelated-cue cases and compared their subsequent revision gains with those of the remaining proposed-method rounds. The outcome was the change in third-party creativity score from the previous round, ΔS_{step}.

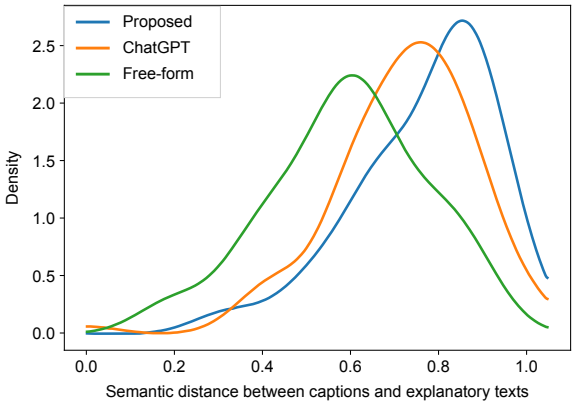


Figure 6: Distributions of semantic distances between captions and explanatory texts in the English domain.

As shown in Figure 7, perceived-unrelated cue rounds showed a negative mean change ($M = -1.20$, $n = 12$), whereas the remaining proposed-method rounds showed a positive mean change ($M = 2.67$, $n = 73$). This pattern suggests that cues perceived as unrelated may hinder revision, while cues that remain meaningfully connected to the caption may support improvement.

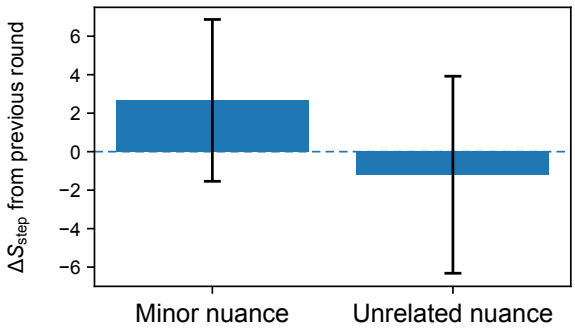


Figure 7: Comparison of revision gains between perceived-unrelated cue rounds and other proposed-method rounds. The vertical axis shows the change in third-party creativity score from the previous round, ΔS_{step}. Error bars indicate 95% confidence intervals.

A.6 Experiment Interface

This appendix shows screenshots of the prototype interface used in the participant experiment (Section 5). Figure 8 illustrates the three main screens used during the image captioning task: the screen for annotating a pair of minor nuances, the baseline revision screen using another participant's free-

Prompt type	Prompt
Judge prompt	You are an expert in nuance associated with words. For each nuance, classify it as either "easy to associate" or "difficult to associate" from the given word and briefly explain your judgment.
Prompt 1	Output 10 nuances that can be associated with the given word, considering minor perspectives, in English phrases of about one clause each.
Prompt 2	Two people are brainstorming ideas to create a creative caption for an image. To help one person understand the nuance of the word proposed by the other, output 10 possible nuances for the word in English, each about one clause long. Include minor perspectives and extreme viewpoints that people would not normally consider.

Table 8: Prompts used in the preliminary LLM comparison.

Table 9: Mean number of difficult-to-associate nuances per input among 10 generated nuances, evaluated by ChatGPT-4o and Claude Fable 5.

Generator	ChatGPT-4o	Claude Fable 5
COMET	4.03	7.37
ChatGPT-4o (Prompt 1)	0.13	0.50
ChatGPT-4o (Prompt 2)	0.20	4.43

form explanation, and the proposed-method revision screen using another participant's annotated minor-nuance pair.

Prompt type	Prompt
Nuance-generation prompt	Given a subjective expression (X), generate two nuance phrases that can be associated with X but are as weakly related to each other as possible. Prioritize minor or extreme perspectives that would not readily come to mind. Do not generate synonyms, near-synonyms, or a hypernym–hyponym pair. Return only valid JSON in the following format: `nuance A:" ", nuance B:" ", rationale A:" ", rationale B:" "`. Each nuance must contain no more than five words. Each rationale must briefly explain why its corresponding nuance was generated and contain no more than 20 words.

Table 10: Prompt used to generate GPT-5 mini nuances for the semantic-distance analysis. Only `nuance A` and `nuance B` were used in the analysis.



(a) Annotating a minor-nuance pair.



(b) Referring to another participant's initial caption and free-form explanation in baseline condition.



(c) Referring to another participant's initial caption and minor nuances in the proposed condition.

Figure 8: Screenshots of the prototype interface used in the experiment. (a) Participants annotated which of two minor nuances better matched their intended meaning by using a slider ranging from 0 to 1. (b) In the baseline condition, participants revised their captions while referring to another participant's initial caption and free-form explanation. (c) In the proposed-method condition, participants revised their captions while referring to another participant's initial caption and the sender's annotation on a minor-nuance pair.