INJECTING PERCEPTUAL FEATURES INTO T5 FOR FIGURATIVE LANGUAGE GENERATION

Wu Yufeng

Department of Linguistics and Translation, City University of Hong Kong, Hong Kong, China

ABSTRACT

Understanding metaphors remains a core challenge for NLP systems, especially when metaphorical meaning depends on perceptual grounding. This paper explores whether injecting perceptual color features into a T5-based language model can enhance metaphor explanation generation. We propose a low-cost, interpretable approach by mapping 12-dimensional color vectors (JzAzBz space) into prefix embeddings that condition the model during fine-tuning. Evaluation on held-out test sets shows that the color-injected model outperforms the text-only baseline in both automatic metrics (BLEU +144%, ROUGE-L F1 +150%) and human ratings of correctness and general quality. However, a significant drop in comprehensiveness is observed, suggesting a trade-off between precision and coverage. Rater agreement analyses reveal high within-item agreement but modest inter-rater consistency, underscoring the subjective difficulty of metaphor evaluation. Our findings demonstrate the utility of perceptual grounding for figurative language generation and offer insights into balancing accuracy and elaboration in metaphor explanation tasks.

KEYWORDS

Metaphor Explanation, Embodied Cognition, Multimodal NLP

1. INTRODUCTION

Metaphor is a pervasive and powerful mechanism in human language, enabling abstract concepts to be understood through more concrete, embodied experiences. Yet for neural language models, understanding and explaining metaphors remains challenging, especially when figurative meaning cannot be inferred purely from textual context. Inspired by cognitive theories of embodied meaning, this study investigates whether injecting perceptual cues, specifically, visual color features, can improve a language model's ability to generate plausible metaphor explanations.

Recent advances in grounded and multimodal NLP suggest that perceptual features (e.g., vision, sensorimotor norms, affect) can help disambiguate figurative expressions and provide cognitively meaningful structure. Color, in particular, encodes rich affective and conceptual associations across cultures and languages. Prior research has shown that color terms, even in metaphorical use, activate visual processing regions in the brain, supporting the idea that metaphor comprehension is not purely symbolic but grounded in perceptual simulation.

To operationalize this perspective in NLP, we propose a simple yet effective method: injecting JzAzBz color features into a T5 model via prefix-tuning. By projecting low-dimensional

David C. Wyld et al. (Eds): CSEN, NLP – 2025
pp. 15-30, 2025. CS & IT – CSCP 2025 DOI: 10.5121/csit.2025.152102

perceptual cues into prefix embeddings, we enable the model to jointly attend to textual and color-derived information during explanation generation. We compare this color-injected system to a text-only baseline across both automatic metrics (BLEU, ROUGE-L F1) and human evaluations of interpretive quality. Our findings demonstrate that perceptual grounding enhances metaphor interpretation but also reveals trade-offs that prompt deeper questions about the role of multimodal priors in generative NLP.

2. LITERATURE REVIEW

2.1. Embodied Cognition Theory

Early theories of meaning in formal semantics and cognitive psychology are grounded in the Symbolic Model of Meaning. This model holds that concepts are represented in the mind as abstract, amodal symbols, detached from perception or bodily experience [1] [2]. Understanding a term such as red involves accessing a symbolic node within a semantic network, not simulating its perceptual qualities. Semantic processing operates through rule-based mechanisms acting on taxonomic structures [3]. Metaphorical language is treated as lexical extension or polysemy. However, such models have been criticized for failing to explain how abstract concepts are acquired or grounded in human experience [4].

As a paradigm shift, Embodied Cognition proposes that cognitive representations are grounded in the body's sensorimotor and affective systems. Meaning is constructed through the reenactment of perceptual, emotional, and motor experiences [5] [6] [7]. This framework underlies two influential theories: Conceptual Metaphor Theory (CMT) and Perceptual Simulation Theory (PST).

CMT, developed by Lakoff and Johnson [8], argues that abstract concepts are structured by metaphorical mappings from concrete bodily experiences. Common metaphors such as 'time is money," 'anger is heat," and 'up is good" show how sensorimotor interactions provide cognitive scaffolding for abstract reasoning. These metaphors are not just rhetorical expressions, but cognitive tools embedded in everyday thought and language.

Building on CMT, Barsalou's Perceptual Simulation Theory [5] and subsequent work [9] propose that conceptual understanding involves simulating sensory and motor experiences associated with the referent. For example, even in metaphorical contexts, the word 'red" can activate visual regions in the brain. Neuroscientific evidence supports this view. Modality-specific regions, such as those related to motion and vision, are activated during language comprehension of sensory-related words [10] [11]. Thus, language understanding is suggested to not be separate from perception but intimately linked through simulation mechanisms.

Situated Conceptualization Theory further refines embodiment by emphasizing that concepts are not static but are dynamically constructed in context. When a concept is activated, it triggers a situated simulation that re-enacts perceptual, motor, and affective experiences relevant to the situation [12]. Therefore, meaning is context-dependent and rooted in prior embodied interactions.

Extending this view, gradient embodiment models propose that concepts vary in the degree of sensorimotor grounding they involve. Connell & Lynott [13] showed that perceptual strength ratings better predict word processing behavior than traditional concreteness measures. Villani et al. [14] demonstrated that even abstract words differ in how much they engage perceptual systems, reinforcing a continuum of embodiment rather than a binary classification.

Complementing embodied theories, Dual Coding Theory [15] proposed that cognition involves two interacting systems, a verbal system for linguistic processing and a non-verbal system for sensory-based representations such as imagery. Words with strong perceptual grounding can activate both systems simultaneously, leading to more robust comprehension and memory. In the context of color metaphors, this dual activation mechanism explains the enhanced processing of abstract terms like ‘red alert” or ‘white lie”, where both symbolic and perceptual codes are engaged.

To reconcile symbolic and embodied perspectives, recent research has proposed hybrid models that integrate both. Andrews et al. [16] and Dove [17] argue that while core semantic structures may be symbolic, they are dynamically enriched by perceptual, affective, and contextual information. For instance, white may function as a symbol of purity in one context while evoking visual experiences of brightness or emptiness in another. These models provide a more flexible account of meaning, bridging compositional semantics with experiential grounding.

Empirical research provides strong evidence for embodied cognition. Neuroimaging studies corroborate this view: language comprehension activates sensory and motor areas, with Pulvermüller [10] reporting motor activation for action words and Simmons et al.[11] showing that even reading color terms activates the brain’s color-processing region.

While early research primarily examined concrete concepts, recent work has extended embodiment to abstract domains such as emotion, morality, and social cognition. Connell et al. [18] showed that abstract concepts vary in their reliance on different perceptual modalities. For example, moral concepts tend to engage motion systems, while numerical concepts often rely on visual simulations. Similarly, Lynott et al. [19] provided large-scale evidence that abstract and concrete words both evoke modality-specific sensory responses, which is consistent with Barsalou [12] view of grounded abstraction. Studies by Villani et al. [14] and Kousta et al. [20] highlight that affective valence, interoception, and social context can serve as grounding mechanisms for abstract concepts when direct perceptual simulation is limited.

Among perceptual domains, color has proven especially useful for testing theories of visual grounding. Simmons et al. [11] found that reading color terms, even metaphorically, activates the brain’s color-processing area, underscoring the embodied nature of color-language links. Lupyan [21] proposed the Label-Feedback Hypothesis, which further suggests that linguistic labels actively shape perceptual expectations. Evidence from cross-linguistic studies [22] and developmental research shows that language can modulate visual grouping.

Building on this foundation, researchers have explored multimodal and image-based approaches to embodiment. Guilbeault et al. [23] demonstrated that abstract concepts are systematically associated with specific visual features in corresponding images, such as hue, brightness, entropy, and shape irregularity. These visual regularities were shown to align with affective and conceptual similarity, offering strong evidence for visual simulation in abstract conceptualization. This image-based embodiment perspective has been extended to Chinese. Hui et al. [24] found that color-emotion congruence facilitated faster semantic judgments among Chinese speakers, suggesting that visual-emotional mappings influence real-time language processing. Their findings further support the idea that perceptual cues like color are not only cognitively active but also culturally embedded.

Recent research in embodied cognition has increasingly highlighted the role of visual information in shaping conceptual understanding. Studies in information visualization have shown that visual representations are not only tools for representing abstract data but also play an active role in how information is perceived, processed, and semantically interpreted [25][26]. Schloss et al. [27]

demonstrated that concept-color mappings are systematically structured in cognition, suggesting that viewers intuitively associate specific hues with abstract ideas like threat or purity. The semantic discriminability of visual stimuli supports the idea that perceptual clarity and feature salience contribute to concept recognition and cognitive efficiency [27]. Interaction-based views reinforce this, emphasizing that 3D visualization environments with embodied controls foster intuitive spatial reasoning by engaging motor schemas [28] [29]. As such, visual semantics are not limited to surface features but involve deeply embodied interpretations grounded in perceptual and interactional experience. The findings across visualization design, perceptual cognition, and metaphor theory underscore the relevance of visual grounding in meaning construction.

2.2. Metaphor Prediction in NLP

Early computational treatments of metaphor framed the task as detection: determining whether a token or phrase is used metaphorically in context. This line of work relied on carefully specified manual protocols such as the Metaphor Identification Procedure (MIP) and its successor MIPVU, which standardized annotation criteria and seeded later supervised models [30] [31]. These protocols remain the basis of widely used corpora and shared tasks.

Benchmark datasets then catalyzed progress. The VU Amsterdam Metaphor Corpus (VUA) underpinned the 2018 shared task, which evaluated systems on word-level metaphor identification across genres; VUA helped establish common train/test splits and metrics and revealed difficulties such as class imbalance and cross-genre robustness [32]. Complementary resources include MOH-X (verb-focused, high metaphor ratio) and TroFi (literal vs. figurative verb usages from WSJ), often used for in-context classification or sequence labeling [33].

Modeling evolved from feature-engineered SVM/CRF baselines to neural architectures. Early neural systems used BiLSTMs over contextual windows or sequence tagging; as pre-trained contextual encoders arrived, BERT-based models became dominant in shared tasks and follow-up studies (e.g., reports noting neural dominance at VUA) [34].

Specialized architectures then embedded metaphor-theoretic biases into transformers. MelBERT, for example, employs a late-interaction mechanism over BERT informed by identification theories, achieving strong performance on VUA, MOH-X, and TroFi [35]. Such designs illustrate how inductive bias plus large pre-training can outperform naive fine-tuning for figurative language.

Parallel to purely textual modeling, researchers showed that non-textual cues can help. Shutova et al. [36] incorporated visual features to detect metaphor, demonstrating that multimodal signals boost accuracy when linguistic context is ambiguous, an insight that motivates structured perceptual cues (e.g., affect or color) as auxiliary inputs even in text-only pipelines.

A second shift reframed tasks from detection to generation: producing literal explanations or paraphrases of metaphorical language. Text-to-text models like T5 unify classification, tagging, and generation by casting problems as string-to-string mapping, enabling “explain the metaphor” prompts and multi-task training within a single architecture [37].

Current challenges include variable inter-annotator agreement (even with MIP/MIPVU), domain shift across genres, and limited coverage for non-English metaphors. Recent work addresses these via better annotation protocols, domain adaptation, and multilingual pre-training; nevertheless, careful evaluation on both automatic metrics and human judgments remains crucial for explanation tasks, not just detection [38].

Finally, there is growing interest in grounded or knowledge-augmented approaches that inject structured cues tied to conceptual mappings (e.g., affect, concreteness, or color priors) or leverage multimodal encoders. These align with cognitive accounts that treat metaphor as systematic mapping rather than noise, and they complement the strong baselines provided by general-purpose text-to-text transformers [36].

2.3. Grounded and Multimodal LLMs

Large language models (LLMs) increasingly go beyond text-only inputs to incorporate grounded signals, perceptual, affective, or world-knowledge cues that humans routinely exploit when interpreting meaning. This move is motivated by embodied cognition: many linguistic meanings, including metaphors, are systematically tied to sensorimotor experience [4] [5]. If abstract language draws on concrete experience, then LLMs should benefit from multimodal inputs (vision, sound, touch) or structured proxies of perception (e.g., color, sensorimotor norms) when tasked with nuanced semantic understanding and figurative interpretation.

Vision as grounding. Early grounding efforts fused text with image features to overcome the limits of purely distributional semantics. Multimodal distributional semantics showed that adding visual descriptors improves concept similarity and lexical inference over text alone [39]. In figurative language, visual features have been used to sharpen metaphor detection, improving disambiguation when context is underspecified [36]. Follow-up work constructed "visibility" or image-derived representations and combined them with textual encoders, yielding gains on benchmark metaphor datasets relative to text-only models [40]. More broadly, modern vision–language systems like CLIP-style pipelines [34] demonstrate how image embeddings can be aligned with text encoders, offering a template for grounded LLM design even when the downstream task is text generation.

Beyond vision: sensorimotor and affective cues. Grounding does not require raw pixels. Human sensorimotor norms, ratings of how words are experienced via sight, sound, touch, taste, smell, or action, provide compact, cognitively meaningful features. Injecting such vectors into neural models improves metaphor identification and yields more interpretable decisions that track human intuitions [41] [42]. Similarly, combining visually grounded word vectors with sensorimotor norms enriches transformer representations on semantic tasks related to figurative meaning [43]. These results suggest that structured, low-dimensional proxies of perception can serve as effective grounding signals without the engineering overhead of full multimodal pretraining.

Color as an embodied prior. Color semantics are psychologically salient and culturally pervasive (e.g., "feeling blue," "in the red"). Large-scale lexicons show stable word–color associations, even for abstract concepts, providing priors that can be turned into features [44] [45]. Perceptually grounded embeddings like comp-syn represent a concept by the distribution of colors in its images, using a uniform color space to produce 8–16 dimensional descriptors [46]. These embeddings complement text: they better predict human concreteness and help separate metaphorical from literal adjective–noun pairs, revealing regularities that text-only models often miss. Color thus offers a lightweight embodied cue well suited to conditioning LLMs for metaphor explanation.

Parameter-efficient grounding for LLMs. Rather than retraining an LLM end-to-end, parameter-efficient methods insert small learnable modules while keeping the backbone largely frozen. Adapters [47], LoRA [48], and prefix/prompt-tuning [49] are especially attractive for integrating grounded signals: a tiny network maps a perceptual vector (e.g., color or sensorimotor features) into continuous prompts/prefixes that the transformer attends to during encoding and generation.

This strategy preserves general linguistic competence while injecting task-relevant grounding.

Proof of concept with "frozen" LMs. Tsimpoukelli [50] showed that a frozen language model can be endowed with visual understanding by learning a vision encoder that outputs a prefix sequence consumable by the LM, enabling few-shot multimodal tasks. Analogously, vision–language variants of T5 (e.g., VL-T5 families) encode images as token-like embeddings fed to the text encoder. These designs validate a general recipe: keep the LM mostly intact, learn a small bridge from perceptual cues to the LM's hidden space, and let attention do the fusion.

Implications for metaphor understanding. Figurative language often hinges on embodied contrasts (temperature, brightness, heaviness, vividness). Grounded LLMs, augmented with visual, sensorimotor, or color priors, are better positioned to choose the intended non-literal sense and to produce literal explanations aligned with human intuitions. In generation, low-dimensional cues stabilize decoding (reducing generic glosses) and bias the model toward interpretations consistent with common embodied mappings.

Positioning of the present study. In this landscape, conditioning a T5 model on compact color vectors via a learned prefix encoder is a principled, low-cost instantiation of grounded LLM design. It operationalizes embodied theory with interpretable, language-agnostic cues, avoids heavy multimodal pretraining, and supports clear ablations (text-only vs. shuffled/zeroed color features). The broader literature on grounded and multimodal LLMs thus motivates and contextualizes our approach to metaphor explanation.

3. METHODOLOGY

3.1. Data

The annotated metaphor data from previous research [51] is used in this research. The dataset is divided into three parts. First, a pretraining corpus of 21,871 metaphorical sentence instances was assembled, in which each instance was paired with an explanatory expression for the target metaphorical word; this corpus was used to induce general mappings from metaphor-in-context to literal explanation. Second, a color-augmented fine-tuning set of 800 instances was curated, where each sentence contained a metaphorical word annotated with perceptual color features and was paired with an explanatory output; this set was used to enable incorporation of visual priors during generation. Finally, two evaluation sets of 200 instances each were prepared: (i) a text-only set, containing sentences with metaphorical words but without color features or gold explanations, by which the baseline model was assessed; and (ii) a multimodal set, containing sentences with metaphorical words and associated color features but no references, by which the color-injected model was assessed.

3.2. Finetuning the Models

An T5-small encoder–decoder model was fine-tuned on a metaphor-explanation dataset to establish a text-only baseline. Training was run for five epochs (batch size = 8; max input length = 128; max output length = 64; learning rate = 5×10^{-5}). The resulting checkpoint (t5_finetuned_metaphor) served as the baseline for subsequent comparisons.

To assess whether color-related perceptual information could enhance explanation quality, a Color Prefix Encoder was introduced to project 12-dimensional JzAzBz color features into a sequence of prefix embeddings. These embeddings were concatenated with the token embeddings at the encoder input so that textual and perceptual cues could be jointly attended. Joint

optimization was performed over both the prefix encoder and the T5 backbone during finetuning (five epochs; batch size = 8; learning rate = 2×10^{-4}). The trained multimodal system (joint_model) and its prefix weights (color_prefix_encoder.pt) were saved as the color-injected model.

3.3. Prediction Generation

For the text-only baseline, held-out metaphorical sentences were tokenized and decoded with greedy or beam search (maximum length = 50), and the outputs were saved to baseline_predictions.csv.
For the color-injected model, the same items were paired with their color feature vectors. Each feature vector was mapped by the Color Prefix Encoder to prefix embeddings, concatenated with token embeddings, and then processed by the backbone to generate explanations. Decoded outputs were saved to color_injected_predictions.csv. This protocol enabled a controlled comparison between text-only and color-informed generations.

3.4. Evaluation of Model Outputs

Automatic evaluation was conducted using sentence-level BLEU (with smoothing for short sequences) and ROUGE-L F1, computed against human reference explanations where available, and then averaged to obtain corpus-level scores for each model.

A complementary human evaluation was implemented using a balanced paired-questionnaire design. The 34 test sentences were split into two halves; in Questionnaire A, baseline outputs for one half and color-injected outputs for the other were presented, with the assignments reversed in Questionnaire B. In this way, coverage of all sentences was ensured while direct A/B exposure per item was avoided. Eligible participants ($\geq$ 18 years) completed exactly one questionnaire ($\approx$ 10–15 minutes) and rated each explanation on correctness, grammaticality, naturalness, and comprehensiveness using a five-point Likert scale. Quality-control items were embedded to safeguard response validity.

Human ratings were summarized with descriptive statistics for each model and dimension. Paired comparisons between baseline and color-injected outputs were performed at the sentence level using paired t-tests and Wilcoxon signed-rank tests; Cohen's d was reported to quantify effect sizes. Rater consistency was examined via Cronbach's α (internal consistency within each rater group), within-group agreement (rwg) computed per item, and quadratic-weighted Cohen's κ across rater pairs. Through these analyses, the automatic metrics were validated with human judgments and the impact of color-feature injection on perceived explanation quality was assessed.

4. RESULTS

We evaluate metaphor explanation on two held-out test sets: a text-only set for assessing the baseline and a color-augmented set for assessing the multimodal system. The compared systems are (i) a Baseline model, T5-small fine-tuned on text only, and (ii) a Color-Injected model, the same backbone jointly fine-tuned with a learned color-prefix that conditions on 12-D color features. We report corpus-level BLEU and ROUGE-L F1 for automatic evaluation, and human ratings on five dimensions (correctness, grammaticality, naturalness, comprehensiveness, general quality) using a 5-point Likert scale with 6 raters × 34 items per form. For significance, sentence-paired t-tests and Wilcoxon signed-rank tests are applied, and Cohen's d is provided to quantify effect size.

4.1. Automatic Metrics

Table 1. Automatic metrics on the metaphor-explanation test set.

Model	BLEU	ROUGE-L (F1)
Baseline (text-only)	0.0018	0.0100
Color-Injected	0.0044	0.0250

The color-injected model outperforms the text-only baseline on both metrics: BLEU +144% (0.0044 vs. 0.0018) and ROUGE-L F1 +150% (0.0250 vs. 0.0100). While absolute scores are low, typical for open-ended explanation generation, the consistent relative gains indicate that conditioning on compact color features helps the model better align with reference explanations under the same decoding setup described in Methods (sentence-level BLEU with smoothing; ROUGE-L F1).

4.2. Human Evaluation

4.2.1. Descriptive Statistics (All Ratings)

Each model/dimension has N = 204 judgments (34 sentences × 6 raters). Means and standard deviations (SD) are summarized below.

Table 2. Summary of human ratings

Dimension	Baseline Mean	Baseline SD	Color-Injected Mean	Color-Injected SD
Comprehensiveness	3.75	1.17	3.17	1.09
Correctness	3.73	1.29	4.01	1
General quality	3.48	1.45	3.82	1.06
Grammaticality	3.62	1.02	3.73	0.81
Naturalness	3.68	1.15	3.67	1.09

These results offer an initial overview of the perceived quality of metaphor explanations. On Correctness, the color-injected model shows a noticeable improvement over the baseline (mean: 4.01 vs. 3.73), suggesting that the injected perceptual cues helped produce more semantically accurate outputs. A similar pattern emerges for General quality, where the multimodal system achieved a higher mean (3.82 vs. 3.48), indicating that raters generally preferred these explanations overall.

For Grammaticality and Naturalness, the two models performed comparably, with slight differences that fall within the range of standard deviation, suggesting that color conditioning did not negatively impact fluency or syntactic well-formedness.

Interestingly, the Comprehensiveness dimension saw a moderate decline in ratings for the color-injected model (3.17 vs. 3.75), indicating that while the generated explanations may have become more accurate or natural, they might have been perceived as less detailed or holistic in covering the metaphor's meaning.

4.2.2. Sentence-Paired Significance Tests

To assess whether differences between the two systems were statistically meaningful, we conducted sentence-level paired significance tests across all five human rating dimensions. Each

of the 34 test sentences received ratings from six participants per model, and mean scores were aggregated per sentence, resulting in 34 paired observations for each model-dimension comparison. Two statistical tests were applied per dimension: (i) paired t-tests to assess mean differences, and (ii) Wilcoxon signed-rank tests for robustness against non-normality. Cohen's d was reported to indicate effect size.

Table 3. Sentence-level paired test results.

Dimension	Mean (Base)	Mean (Fine)	Δ (Fine–Base)	t	p (t)	Wilcoxon W	p (W)	Cohen's d
Comprehensiveness	3.75	3.17	−0.58	−5.10	0.00	55.00	0.00	−0.87
Correctness	3.73	4.02	0.28	2.73	0.01	131.50	0.01	0.47
General quality	3.48	3.82	0.35	3.40	0.00	125.00	0.00	0.58
Grammaticality	3.62	3.73	0.11	1.28	0.21	225.00	0.26	0.22
Naturalness	3.68	3.67	−0.02	−0.14	0.89	290.00	0.91	−0.02

The color-injected model significantly outperformed the baseline on Correctness ($p < 0.01$, $d = 0.47$) and General quality ($p < 0.005$, $d = 0.58$), with medium effect sizes. These results suggest that perceptual color cues helped the model choose more appropriate interpretations for metaphorical expressions and improved the overall perceived quality of explanations.

No statistically significant differences were found in Grammaticality or Naturalness, with small effect sizes ($|d| < 0.25$). This indicates that the incorporation of color features did not compromise fluency or syntactic well-formedness, maintaining linguistic plausibility at a level similar to the text-only baseline.

In contrast, the model showed a significant decline in Comprehensiveness ($p < 0.001$, $d = -0.87$), with a large negative effect size. This suggests that although the color-injected model generated more accurate and well-received explanations, these outputs may have covered a narrower scope of meaning or omitted details, likely due to the model's stronger anchoring on the dominant perceptual cue (color).

Together, these findings highlight a tradeoff: color conditioning boosts interpretive precision and global acceptability but may reduce the breadth of the explanation. This mirrors prior findings in multimodal NLP where perceptual features strengthen grounding but may also bias attention toward more concrete aspects of meaning.

4.2.3. Rater Reliability and Agreement

Rater consistency was examined with three complementary indicators, Cronbach's α (internal consistency within each six-rater group), r_wg (within-group agreement per item under a uniform 5-point null), and quadratic-weighted Cohen's κ averaged across all rater pairs. Because the study used two balanced forms, results are reported separately for Questionnaire A and Questionnaire B to respect the split-by-side design.

(a) Questionnaire A

Table 4. Reliability of questionnaire A.

Dimension	α Base	α Fine	r_wg Mean Base	r_wg Mean Fine
Comprehensiveness	0.48	0.49	0.54	0.41
Correctness	−0.40	−0.22	0.29	0.2
General quality	−0.50	−0.16	−0.05	0.14
Grammaticality	−0.44	−0.57	0.58	0.5
Naturalness	−0.07	−0.08	0.46	0.4

Table 3A (above) summarizes three complementary indices: Cronbach's α (internal consistency across the six raters), r_wg (within-group agreement per item under a uniform 5-point null; higher is better), and the mean quadratic-weighted Cohen's κ across all rater pairs (agreement beyond chance).

First, comprehensiveness shows the most stable internal consistency in Form A: α ≈ 0.48 (Baseline) and 0.49 (Color-Injected), i.e., approaching the conventional "moderate" region for small panels. At the same time, within-item agreement (r_wg) is higher for Baseline (0.543) than for Color-Injected (0.409). This pattern suggests that the six raters behaved relatively coherently as a group across items (reasonable α) and, on average, converged more tightly on the Baseline outputs than on the Color-Injected ones for this scale.

Correctness exhibits low consensus: α is negative for both systems (−0.400, −0.223), and r_wg is modest (0.286, 0.199). Negative α signals weak or inconsistent co-movement among raters across items, often a sign that raters applied heterogeneous criteria (e.g., emphasis on literal fidelity vs. interpretive plausibility) or that the construct blends subfacets. The low r_wg mirrors this, indicating notable dispersion within items.

General quality shows the starkest divergence: α is negative in both systems (−0.503, −0.162), and the Baseline even has a slightly negative mean r_wg (−0.051), meaning dispersion sometimes exceeded what would be expected by chance on a 5-point scale. The Color-Injected condition rises to a small positive r_wg (0.141), but agreement remains weak. This is typical for broad "overall" scales that bundle multiple cues (correctness, coverage, fluency), inviting rater idiosyncrasies.

Grammaticality achieves the strongest within-item agreement of the five dimensions (r_wg ≈ 0.58 Baseline; 0.50 Color-Injected), consistent with raters more easily aligning on surface well-formedness. Yet α remains negative in both systems (−0.436, −0.570), implying that, although raters tend to agree within each item, they do not track each other consistently across different items (e.g., a rater can be strict on one subset and lenient on another in ways that do not align with peers). This "high r_wg but low α" pattern occurs when item-to-item rater rank orders are unstable.

Naturalness falls between these extremes: r_wg is moderate (0.455, 0.402), but α hovers near zero (−0.069, −0.077). As with grammaticality, raters can often align within a given item on how "native-like" an explanation feels, yet their across-item rating profiles diverge.

Finally, κ (weighted, averaged over all rater pairs) sits near zero across dimensions in Form A (e.g., comprehensiveness 0.097 for Baseline, 0.114 for Color-Injected; other dimensions near or slightly below zero). This indicates that, despite acceptable within-item agreement on some dimensions (especially grammaticality and sometimes comprehensiveness), individual rater pairs did not exhibit strong beyond-chance alignment on absolute category choices. In practice, that

means group means are usable (supported by r_wg), but any single rater's labels are noisy relative to another's.

(b) Questionnaire B

Table 5. Reliability of questionnaire B.

Dimension	α Base	α Fine	r_wg Mean Base	r_wg Mean Fine
Comprehensiveness	−0.01	−0.04	0.39	0.53
Correctness	−0.26	0.27	0.61	0.53
General quality	0.19	−0.45	0.52	0.50
Grammaticality	−0.20	−0.22	0.72	0.72
Naturalness	−0.43	−0.01	0.47	0.51

Comprehensiveness shows a notable shift compared with Form A: within-group agreement (r_wg) is higher for the color-injected system (0.531) than for the baseline (0.390), suggesting raters converged more on how fully the color-conditioned outputs covered the intended meaning. Cronbach's α remains near zero for both systems (−0.008, −0.037), pointing to weak inter-rater covariance across items, i.e., raters' profiles vary from item to item even when they agree within a given item.

Correctness exhibits the clearest reliability gain for the color-injected system in Form B. Internal consistency rises to a small positive α (0.271) from a negative baseline (−0.262), and r_wg remains strong for both systems (0.614 baseline; 0.533 color-injected). This pattern suggests that while both systems produced outputs on which raters could agree per item, the color-injected outputs yielded more consistent rater rank-ordering across items (higher α), possibly because color cues stabilized judgments of literal/interpretive fidelity.

General quality presents mixed evidence. Within-group agreement is solid and similar across systems (r_wg ≈ 0.51 vs. 0.50), yet α is positive for the baseline (0.193) and negative for the color-injected system (−0.447). This indicates that, although raters converged within items for both systems, their across-item co-movement diverged more under the color-injected condition, consistent with "overall quality" being a composite construct where individual raters emphasize different facets (coverage, fluency, faithfulness).

Grammaticality again shows the strongest within-item consensus of all dimensions (r_wg ≈ 0.72 for both systems), reflecting that surface well-formedness is relatively easy to align on. Yet α remains slightly negative for both, implying that raters' strictness varies across items in ways that do not tightly track one another (high r_wg, low α).

Naturalness achieves moderate r_wg for both systems, with a small advantage for color-injected (0.508 vs. 0.471). α improves from strongly negative for baseline (−0.430) to near zero for color-injected (−0.013), again hinting that color conditioning may reduce idiosyncratic variation in how "native-like" an explanation feels across items.

Finally, the consistently near-zero κ values confirm that individual rater pairs did not agree far beyond chance on absolute category choices, even when r_wg indicated good within-group convergence. Practically, this supports our use of aggregated per-item means and sentence-paired tests (rather than single-rater decisions) and motivates two follow-ups for future data collection: provide tighter rubric anchors and short calibration examples per dimension, and consider many-

facet models (e.g., Rasch/GLMM) to absorb rater-severity differences.

5. DISCUSSION

This study explored whether injecting perceptual color features into a T5 model improves performance on the metaphor explanation task. Automatic metrics showed consistent gains: the color-injected model outperformed the baseline with over 2× improvement in both BLEU and ROUGE-L F1. Human evaluations echoed these gains, with color-injected outputs rated significantly higher in correctness and general quality, and exhibiting moderate effect sizes (Cohen's $d \approx 0.5$). However, a notable trade-off emerged: comprehensiveness ratings dropped significantly under color conditioning. Grammaticality and naturalness showed no reliable differences.

5.1. Interpretation and Implications

The improvements in correctness and general quality suggest that color features help steer the model toward more accurate and coherent interpretations of metaphorical meanings. This aligns with the hypothesis that perceptual grounding, in this case via color priors, can aid in metaphor understanding by reinforcing likely conceptual mappings. The moderate effect sizes observed in human ratings, despite small training sets and low-dimensional features, highlight the promise of low-cost, interpretable multimodal cues for generation tasks.

The comprehensiveness drop is revealing. It suggests that while color features may anchor the model's attention to salient conceptual associations, they may also constrain the breadth of interpretation, leading to narrower, more specific (but less elaborative) explanations. This finding resonates with cognitive theories of embodiment: grounding can enhance relevance but reduce generality. In generative terms, conditioning on perceptual priors may increase precision at the expense of recall over interpretive possibilities. Moreover, interpreting these trade-offs is complicated by variability in human evaluation. Although within-item agreement was moderate, overall rater consistency was limited. Future work should therefore include improved rater training protocols or a larger rater pool to strengthen the robustness of human evaluation.

5.2. Rater Behaviour and Evaluation Methodology

Rater agreement patterns help contextualize the results. Across both questionnaires, r_wg values (within-item agreement) were moderate-to-high for dimensions like grammaticality and correctness, supporting the reliability of mean scores. However, Cronbach's α was often near or below zero, indicating inconsistent rank ordering across items, and Cohen's κ hovered near chance. These patterns suggest that while raters can converge on judgments within a sentence, their overall rating styles vary widely.

This heterogeneity reflects the difficulty of the task: evaluating metaphor explanations involves balancing multiple criteria (interpretive fidelity, elaboration, fluency), and individual raters likely weigh these differently. The findings underscore the need for improved annotation protocols in figurative language evaluation, such as rubric anchoring, calibration items, or facet-based rating models.

5.3. Limitations

Several limitations temper the findings. First, the training and test sets are relatively small, which may limit generalizability. Second, the perceptual grounding is based solely on color features; other embodied cues (e.g., sensorimotor, affective, situational) remain unexplored in this setup. Third, while the prefix-injection mechanism is interpretable and efficient, it introduces additional parameters and training complexity that may not scale well to larger models or more diverse metaphors. Finally, our human evaluation relied on a relatively small panel of six raters without extensive calibration. Although within-item agreement was moderate, future work should incorporate more rigorous rater training or a larger pool of annotators to improve consistency and reliability of judgments.

5.4. Future Directions

Future work can expand in several directions. One avenue is to scale up multimodal metaphor datasets, incorporating a wider range of grounded features (e.g., sensorimotor norms, visual entropy, emotion associations). Another is to compare alternative grounding methods, such as image-conditioned decoders or attention-based fusion. From an evaluation perspective, it would be fruitful to combine human ratings with retrieval-based or discriminative probes, testing whether grounded explanations align more closely with human interpretations across diverse metaphor types and domains. Finally, the broader implications of grounding for generative controllability and explainability in LLMs merit systematic exploration.

6. CONCLUSION

This study presents a parameter-efficient method for injecting perceptual color features into a T5 model to improve metaphor explanation. The results show that color-injected models achieve meaningful gains in both automatic and human evaluation metrics, particularly in correctness and general quality. These improvements suggest that perceptual grounding, here instantiated via compact color vectors, can steer generative models toward more accurate interpretations of metaphorical language.

At the same time, the observed drop in comprehensiveness points to a trade-off: color cues help constrain interpretations but may limit elaboration. This finding aligns with theories of embodied cognition that highlight both the benefits and limitations of grounding in perceptual experience. Our rater agreement analysis further reveals the complexity of evaluating figurative explanations, underscoring the need for more structured annotation protocols.

Looking forward, this work lays a foundation for broader integration of grounded signals into LLMs, beyond color to affect, sensorimotor norms, and visual semantics. It also raises important questions about the balance between precision and generative diversity in explainable NLP. By embedding perceptual priors into text-generation pipelines, we take a step toward cognitively plausible models of figurative understanding.

REFERENCES

[1] Fodor, J. A. (1975). The language of thought. Harvard university press Cambridge, MA.
[2] Pylyshyn, Z. W. (1984). Computation and cognition.
[3] Collins, A. M., & Quillian, M. R. (1969). Retrieval time from semantic memory. Journal of Verbal Learning and Verbal Behavior, 8(2), 240–247.
[4] Lakoff, G., Johnson, M., & Sowa, J. F. (1999). Review of Philosophy in the Flesh: The embodied mind and its challenge to Western thought. Computational Linguistics, 25(4), 631–634.
[5] Barsalou, L. W. (1999). Perceptual symbol systems. Behavioral and Brain Sciences, 22(4), 577–660.
[6] Gallese, V., & Lakoff, G. (2005). The brain's concepts: The role of the sensory-motor system in conceptual knowledge. Cognitive Neuropsychology, 22(3-4), 455–479.
[7] Glenberg, A. M., & Kaschak, M. P. (2002). Grounding language in action. Psychonomic Bulletin & Review, 9(3), 558–565.
[8] Lakoff, G., & Johnson, M. (1980). The metaphorical structure of the human conceptual system. Cognitive Science, 4(2), 195–208.
[9] Gibbs Jr, R. W. (2005). Embodiment and cognitive science. Cambridge University Press.
[10] Pulvermüller, F. (2005). Brain mechanisms linking language and action. Nature Reviews Neuroscience, 6(7), 576–582.
[11] Simmons, W. K., Ramjee, V., Beauchamp, M. S., McRae, K., Martin, A., & Barsalou, L. W. (2007). A common neural substrate for perceiving and knowing about color. Neuropsychologia, 45(12), 2802–2810.
[12] Barsalou, L. W. (2010). Grounded cognition: Past, present, and future. Topics in Cognitive Science, 2(4), 716–724.
[13] Connell, L., & Lynott, D. (2012). Strength of perceptual experience predicts word processing performance better than concreteness or imageability. Cognition, 125(3), 452–465.
[14] Villani, C., Lugli, L., Liuzza, M. T., Nicoletti, R., & Borghi, A. M. (2021). Sensorimotor and interoceptive dimensions in concrete and abstract concepts. Journal of Memory and Language, 116, 104173.
[15] Paivio, A. (1991). Dual coding theory: Retrospect and current status. Canadian Journal of Psychology/Revue Canadienne De Psychologie, 45(3), 255.
[16] Andrews, M., Frank, S., & Vigliocco, G. (2014). Reconciling embodied and distributional accounts of meaning in language. Topics in Cognitive Science, 6(3), 359–370.
[17] Dove, G. (2016). Three symbol ungrounding problems: Abstract concepts and the future of embodied cognition. Psychonomic Bulletin & Review, 23(4), 1109–1121.
[18] Connell, L., Lynott, D., & Banks, B. (2018). Interoception: The forgotten modality in perceptual grounding of abstract and concrete concepts. Philosophical Transactions of the Royal Society B: Biological Sciences, 373(1752), 20170143.
[19] Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020a). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. Behavior Research Methods, 52(3), 1271–1291.
[20] Kousta, S., Vigliocco, G., Vinson, D. P., Andrews, M., & Del Campo, E. (2011). The representation of abstract words: why emotion matters. Journal of Experimental Psychology: General, 140(1), 14.
[21] Lupyan, G. (2012). Linguistically modulated perception and cognition: The label-feedback hypothesis. Frontiers in Psychology, 3, 54.
[22] Thierry G, Athanasopoulos P, Wiggett A, Dering B, Kuipers JR. Unconscious effects of language-specific terminology on preattentive color perception. Proc Natl Acad Sci U S A. 2009 Mar 17;106(11):4567-70.
[23] Guilbeault, D., Nadler, E. O., Chu, M., Sardo, D. R. L., Kar, A. A., & Desikan, B. S. (2020). Color associations in abstract semantic domains. Cognition, 201, 104306.
[24] Hui, Q., Kong, F., Lin, S., Li, Y., & You, X. (2024). Can orange colour facilitate the processing of happiness? An exploration study on happiness metaphor. International Journal of Psychology, 59(1), 111–120.
[25] Card, S. (2009). Information visualization. Human-computer interaction (pp. 199–234). CRC press.
[26] Chen, C. (2010). Information visualization. Wiley Interdisciplinary Reviews: Computational Statistics, 2(4), 387–403.
[27] Schloss, K. B., Leggon, Z., & Lessard, L. (2020). Semantic discriminability for visual communication. IEEE Transactions on Visualization and Computer Graphics, 27(2), 1022–1031.

[28] Gershon, N., Eick, S. G., & Card, S. (1998). Information visualization. Interactions, 5(2), 9–15.
[29] Kosara, R., Hauser, H., & Gresh, D. L. (2003). An interaction view on information visualization. Eurographics (State of the Art Reports), , 1–5.
[30] Group, P. (2007). MIP: A method for identifying metaphorically used words in discourse. Metaphor and Symbol, 22(1), 1–39.
[31] Nacey, S. (2013). 3. Introduction to MIP (VU). Metaphors in Learner English (pp. 65–80). John Benjamins Publishing Company.
[32] Leong, C. W., Klebanov, B. B., & Shutova, E. (2018a). A report on the 2018 VUA metaphor detection shared task. Paper presented at the Proceedings of the Workshop on Figurative Language Processing, 56–66.
[33] Gao, G., Choi, E., Choi, Y., & Zettlemoyer, L. (2018). Neural metaphor detection in context. arXiv Preprint arXiv:1808.09653,
[34] Gao, Y., Liu, J., Xu, Z., Wu, T., Zhang, E., Li, K., Yang, J., Liu, W., & Sun, X. (2024). Softclip: Softer cross-modal alignment makes clip stronger. Paper presented at the Proceedings of the AAAI Conference on Artificial Intelligence, , 38(3) 1860–1868.
[35] Choi, M., Lee, S., Choi, E., Park, H., Lee, J., Lee, D., & Lee, J. (2021). MelBERT: Metaphor detection via contextualized late interaction using metaphorical identification theories. arXiv Preprint arXiv:2104.13615,
[36] Shutova, E., Kiela, D., & Maillard, J. (2016). Black holes and white rabbits: Metaphor identification with visual features. Paper presented at the Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 160–170.
[37] Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., & Liu, P. J. (2020). Exploring the limits of transfer learning with a unified text-to-text transformer. Journal of Machine Learning Research, 21(140), 1–67.
[38] Leong, C. W., Klebanov, B. B., & Shutova, E. (2018b). A report on the 2018 VUA metaphor detection shared task. Paper presented at the Proceedings of the Workshop on Figurative Language Processing, 56–66.
[39] Bruni, E., Tran, N., & Baroni, M. (2014). Multimodal distributional semantics. Journal of Artificial Intelligence Research, 49, 1–47.
[40] Kehat, G., & Pustejovsky, J. (2020). Improving neural metaphor detection with visual datasets. Paper presented at the Proceedings of the Twelfth Language Resources and Evaluation Conference, 5928–5933.
[41] Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020b). The Lancaster Sensorimotor Norms: multidimensional measures of perceptual and action strength for 40,000 English words. Behavior Research Methods, 52(3), 1271–1291.
[42] Wan, M., Su, Q., Ahrens, K., & Huang, C. (2024). Perceptional and actional enrichment for metaphor detection with sensorimotor norms. Natural Language Engineering, 30(6), 1181–1209.
[43] Kennington, C. (2021). Enriching language models with visually-grounded word vectors and the Lancaster sensorimotor norms. Paper presented at the Proceedings of the 25th Conference on Computational Natural Language Learning, 148–157.
[44] Mohammad, S. (2013). Colourful language: Measuring word-colour associations. arXiv Preprint arXiv:1309.5942,
[45] Mohammad, S. M. (2013). Even the abstract have colour: consensus in word-colour associations. arXiv Preprint arXiv:1309.5391,
[46] Desikan, B. S., Hull, T., Nadler, E. O., Guilbeault, D., Kar, A. A., Chu, M., & Sardo, D. R. L. (2020). comp-syn: Perceptually grounded word embeddings with color. arXiv Preprint arXiv:2010.04292,
[47] Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., & Gelly, S. (2019). Parameter-efficient transfer learning for NLP. Paper presented at the International Conference on Machine Learning, 2790–2799.
[48] Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., & Chen, W. (2022). Lora: Low-rank adaptation of large language models. ICLR, 1(2), 3.
[49] Li, X. L., & Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. arXiv Preprint arXiv:2101.00190,

[50] Tsimpoukelli, M., Menick, J. L., Cabi, S., Eslami, S. M., Vinyals, O., & Hill, F. (2021). Multimodal few-shot learning with frozen language models. Advances in Neural Information Processing Systems, 34, 200–212.

[51] Shao, Y., Yao, X., Qu, X., Lin, C., Wang, S., Huang, S. W., Zhang, G., & Fu, J. (2024). CMDAG: A Chinese metaphor dataset with annotated grounds as CoT for boosting metaphor generation. In Proceedings of the LREC-COLING 2024.

AUTHOR

WU Yufeng is a Ph.D. student in Linguistics and Translation Department of City University of Hong Kong. His research focuses on cognitive linguistics and embodied LLMs.