International Journal of Language and Linguistics
2026, Vol. 14, No. 2, pp. 72–85
https://doi.org/10.11648/j.ijll.20261402.12

SciencePG
Science Publishing Group

Research Article

Performances of LLMs in Multimodal Metaphor Understanding, Generation, Consistency and Creativity Based on FDPEF

Zhong Yuke*

Foreign Studies College, Northeastern University, Shenyang, China

Abstract

Human communication uses the synergistic interaction of multimodalities to express emotions and convey information in this age of rapid information science progress. More scholarly interest in multimodal research has also been sparked by the promotion of multimodal interaction methods; multimodal metaphor research is a novel line of inquiry that emerged from the fusion of interdisciplinary and multimodal discourse research. This study addresses the lack of systematic evaluation of large language models (LLMs) in understanding and generating multimodal metaphors by proposing a four-dimension progressive evaluation framework (FDPEF), based on cognitive linguistics theory and multimodal mechanism. The results indicate that Claude-3-5 leads in understanding ability while Cici is the weakest due to over-abstraction; in terms of generative ability, ChatGPT-4 demonstrates the optimal multimodal mapping logic, but none of the models can completely avoid the "graphic semantic deviation" problem; in terms of consistency, ChatGPT-4 is close to the human-level metaphor comprehension threshold, but still suffers from cognitive bias; and in terms of creativity, LLMs generally rely on the conventional metaphor paradigm, and their creativity is limited by the inherent cognitive framework of the training data. The study shows that LLMs can improve metaphor parsing accuracy through visual-textual joint representation, and can quantify metaphor parsing outcomes and their interpretive transformations into measurable metrics, while its metaphor generation is still limited by path dependence and insufficient understanding of cultural contexts, and needs to be optimized for metaphor controllability in the future by combining multimodal embedding and interpretable AI techniques.

Keywords

Large Language Model, Multimodal Metaphor Performance, Quantitative Evaluation,
Four-dimension Progressive Evaluation Framework (FDPEF)

1. Introduction

In contemporary society characterized by advanced digital communication, human meaning-making has transcended the limitations of a single linguistic modality, shifting towards the synergistic interaction of images, sounds, and texts [8]. Multimodal Discourse Analysis (MDA) is crucial for decoding this complex interaction process by revealing how non-linguistic symbols co-construct meaning alongside linguistic

*Correspondence: Zhong Yuke (775569031@qq.com)

Received: 20 March 2026; **Accepted:** 7 April 2026; **Published:** 24 April 2026

systems [26, 32]. Forceville (1996), in his foundational work Pictorial Metaphor in Advertising, pointed out that multimodal metaphors reconstruct cognitive frameworks through cross-sensory mappings, offering explanatory power far beyond traditional unimodal analysis paradigms [11]. This theory provides essential tools for optimizing communication effectiveness in advertising, film, and human-computer interaction.

Lakoff and Johnson (1980), in Metaphors We Live By, systematically argued that metaphors construct our experiential world through "conceptual mappings" (e.g., "TIME IS MONEY"), deeply shaping human understanding of abstract categories. Metaphors are not merely rhetorical embellishments but fundamental cognitive mechanisms for abstract thinking. In multimedia contexts, multimodal representations of metaphors (such as visual and auditory metaphors) can activate cognitive schemas more efficiently [12], but the complexity of generating and recognizing multimodal metaphors poses significant challenges to existing analytical models.

Despite the demonstrated multimodal content generation capabilities of LLMs like GPT-4, the logical coherence of their metaphor construction and the cognitive alignment of their multimodal mappings remain under-evaluated. Existing research often focuses narrowly on unimodal metaphor understanding tasks (e.g., textual metaphor detection), failing to fully explore the mechanisms of multimodal metaphor generation and the consistency and creativity of model outputs [23].

This study focuses on four core dimensions: (1) multimodal metaphor understanding accuracy (precision in capturing metaphorical intent across modalities), (2) generation capability (rationality and creativity of source-target domain mappings), (3) consistency (self-coherence of understanding and generation logic), and (4) creativity (potential to break conventional metaphor paradigms). By constructing a quantitative evaluation framework that integrates cognitive linguistic rules with generative adversarial validation, this study aims to delineate the cognitive boundaries of LLMs in multimodal metaphor tasks, providing theoretical support for optimizing explainable AI models and intelligent creative systems.

2. Literature Review

2.1. Large Language Models

2.1.1. Development and Application of Large Language Modelling Techniques

In recent years, with the breakthroughs in deep learning and neural network technologies, large language models (LLMs) have made significant progress in the field of natural language understanding and generation, which has pushed forward the development of computational linguistics and artificial intelligence [9]. Representative models such as BERT, T5, and GPT series not only possess powerful semantic analysis capabilities, but also show significant improvement in logical reasoning and generative capabilities [6, 20]. In particular, generative AI such as ChatGPT, with its generalization and intelligent performance, has triggered a technological revolution and gradually transformed into social change [5]. However, these models have also revealed many problems in the process of technological development, such as biased output, high computational cost, lack of real-time autonomous learning capability, and limitations of unimodal interaction [28].

Current LLMs mainly support textual modality interaction, while real-world cognition and understanding require collaborative processing across multiple modalities such as visual, auditory, and linguistic [19]. How to draw on the multimodal processing characteristics of the human brain to realize the integration of multimodal information is the key to improving the general understanding ability of the model. The development of multimodal models can not only make up for the limitations of unimodal language models, but also provide new possibilities for the processing of complex linguistic phenomena such as metaphors. Metaphor, as an important carrier of language and cognition, often involves the integration of multimodal information and the extraction of deeper semantics, so the quantitative assessment of multimodal metaphors has become an important direction for exploring the capabilities of large language models.

2.1.2. Metaphor Understanding and Generation Capabilities of Large Language Models

The ability to process metaphors is an important indicator of the semantic understanding and generation capability of a language model. Studies have shown that large language models can provide support for metaphor understanding and generation by means of semantic analysis and generative probability [22]. For example, the GPT family of models can provide more objective shared semantic information for metaphor processing in addition to semantic analysis and translation relations, thus enabling multidimensional language comparison and deep metaphor modelling [10]. However, the current models are still deficient in terms of consistency and creativity in metaphor processing, especially as the processing of multimodal metaphors needs to rely on the collaborative training and knowledge invocation of multimodal data [19].

2.2. Multimodal Metaphor

2.2.1. Theory and Construction Mechanisms of Multimodal Metaphors

Multimodal metaphor research originated from conceptual metaphor theory and gradually expanded to an interdisciplinary platform. scholars such as Forceville expanded metaphor research to a multimodal level, emphasising the value and significance of modal interaction [27]. This interdisciplinary research not only enhances the breadth of metaphor research, but also promotes the improvement of conceptual metaphor

theory [33]. Multimodal metaphors are able to convey information more intuitively and profoundly with the synergistic effect of multiple modalities such as visual, auditory, and verbal [27].

The construction of multimodal metaphors involves the interaction and integration between multiple modalities. Visual grammar provides an important framework for analysing multimodal metaphors, and by analysing the interaction between visual and linguistic elements, the operation mechanism and meaning expression of metaphors can be revealed [24]. Conceptual integration theory provides a new way of analysing the dynamic narrative of multimodal metaphors and makes up for the inadequacy of traditional metaphor theory in the interpretation of nascent meaning [25].

2.2.2. Application Scenarios and Discourse Analysis

Multimodal metaphors show unique ways of construction and function in different contexts. For example, the construction of multimodal metaphors in children's picture books helps to highlight discourse themes and reveal the functional meaning of metaphors through visual grammar [24]. In political cartoons and advertisements, multimodal metaphors convey cultural connotations and communicative intentions through the combination of visual and verbal language, enhancing the infectious and persuasive power of the message [26]. Multimodal metaphors in poster discourse, on the other hand, enrich the types of metaphorical representations through the combination of different modalities, providing a thinking reference for creative practice [29].

2.3. Multimodal Metaphor Performance of Large Language Models

In recent years, with the development of Large Language Models (LLMs), some initial studies have begun to explore metaphor competence of LLMs.

2.3.1. Multimodal Metaphor Understanding of LLMs

According to Conceptual Metaphor Theory (CMT), metaphors map a target domain onto a source domain, and understanding this mapping becomes a basic need that can capture the nature of metaphors, so a model which can simulate the human cognitive process for identifying mappings is created, named Chain-of-Thought Prompting-based Metaphor Mapping Identification Model [30].

CMT also leverages metaphorical mappings to structure abstract reasoning, so a framework, CMT-based prompts, which contains benchmarks like metaphor identification and mapping, domain-specific reasoning, explanation and teaching tasks, and reading comprehension of metaphors, is formed, comparing four native models (Llama3.2, Phi3, Gemma2, and Mistral) and their CMT-augmented counterparts, demonstrating that metaphorical reasoning can be improved by structuring prompts according to conceptual mapping rules [17].

Despite these advances, most studies focusing on multimodal metaphor understanding remain limited in two key aspects.

First, existing work primarily emphasizes textual metaphor detection or general vision–language alignment, rather than metaphor-specific mapping identification [31]. Although some research explores information flow between language and vision, it rarely targets metaphorical coherence or cognitive consistency at a deep level.

Second, while GPT-4 has shown emergent ability in interpreting novel literary metaphors [14], most existing evaluations rely on surface-level features (e.g., lexical overlap, sentence length) rather than metaphorical content itself [1]. This indicates a critical gap: current LLMs lack robust, cognitively validated mechanisms for understanding the structural logic of multimodal metaphors.

2.3.2. Multimodal Metaphor Generation of LLMs

In the domain of metaphor generation, early attempts have explored the construction of visual metaphors from linguistic inputs by combining LLMs with diffusion models [4]. The HAIVMet framework and dataset demonstrate that LLM–diffusion collaboration can generate visually coherent metaphors under CoT prompting and human curation, providing a promising direction for multimodal figurative language processing [4].

However, two major limitations persist in current generation research. First, most studies focus on text-to-image mapping rather than multimodal metaphor reasoning. That is, they generate visual content based on literal descriptions but fail to ensure that the generated metaphors maintain consistent conceptual mappings with the source text. Second, there is no systematic assessment of generation quality dimensions such as creativity, cultural adaptability, or cognitive consistency. Existing evaluations prioritize visual alignment over metaphorical coherence, making it difficult to determine whether generated metaphors truly reflect abstract conceptual structures [12].

Furthermore, although developing multimodal LLMs has become a key direction for enhancing general perception and cross-modal integration [19], the specific challenges of metaphor generation—such as path dependence, conventional metaphor paradigms, and semantic deviation—remain underexplored [4, 7].

2.3.3. Research Gaps and Motivations

Synthesizing the above literature reveals three critical research gaps that motivate the present study:

First, lack of systematic evaluation frameworks for multimodal metaphor performance. Existing research focuses on isolated tasks rather than a unified framework covering understanding, generation, consistency, and creativity.

Second, insufficient cognitive depth in metaphor pro-

cessing. Most studies address surface-level alignment but neglect conceptual mapping logic, cultural context, and cognitive consistency.

Third, absence of multimodal-specific benchmarking for metaphorical tasks. No standardized protocol exists for measuring LLM performance in complex, cognitively driven multimodal metaphor tasks.

These gaps highlight the necessity of the proposed Four-Dimensional Progressive Evaluation Framework (FDPEF), which integrates cognitive linguistic principles with multimodal mechanisms to provide a comprehensive, interpretable, and systematic assessment of LLM metaphor performance.

3. Methodology

3.1. Research Questions

This study adopted a four-dimension progressive evaluation framework (FDPEF) covering four core dimensions, namely, Understanding ability (UI), Generation ability (GI), Consistency (CI), and Creativity (CrI). In addition, this study conducted a quantitative index (e.g., semantic understanding accuracy, modal coordination, structural-semantic-cognitive congruence, and conceptual span) of the Cici, Kimi, GPT-4, and Claude-3-5, the mainstream models to develop empirical analysis. The research questions, based on the four-dimension progressive evaluation framework, are as follows:

1) What are the performances of four large language models in multimodal metaphor understanding?
2) What are the performances of four large language models in multimodal metaphor generation?
3) What are the performances of four large language models in multimodal metaphor consistency?
4) What are the performances of four large language models in multimodal metaphor creativity?

3.2. Context and Large Language Models

3.2.1. Data Sources and Corpus Selection

The research corpus is selected from the promotional posters of IKEA and WWF. These two types of posters have the following characteristics: IKEA promotional posters: with home design as the theme, the combination of visual and linguistic elements highlights creativity and functionality, and metaphors usually involve the concepts of space, comfort and lifestyle; WWF publicity posters: with environmental protection as the core theme, the interaction between visual and verbal modalities emphasizes emotional impact and social responsibility, and the metaphors are mostly related to nature protection, ecological crisis and human responsibility.

The multimodal metaphor design of these posters is both visually appealing and contains deep-seated semantic expressions, providing an ideal analytical object for the study of the construction mechanism of multimodal metaphors.

3.2.2. Subjects and Model Selection

In this study, four subjects are selected for analysis and comparison, including:

Cici: as a neural large language model from ByteDance, Cici is mainly used for the basic analysis of metaphor understanding, providing stable semantic parsing results.

Kimi: as an emerging language generation model, Kimi has some creativity in metaphor generation, but its ability to process multimodal information is limited.

GPT-4: a generative AI model developed by OpenAI, it has strong metaphor understanding and generation ability, supports multi-round interaction, and can generate complex metaphors in combination with context.

Claude-3.5: a language model developed by Anthropic, focusing on security and alignment, showing high consistency and logic in metaphor generation.

These five model choices cover traditional language processing tools, emerging language generation models and image generation models, providing comprehensive technical support for studying multimodal understanding and generation of multimodal metaphors.

3.3. Four-Dimension Progressive Evaluation Framework (FDPEF)

In the evolving field of natural language processing, the ability to understand and generate metaphors is a critical benchmark for assessing the competence of large language models (LLMs). This study introduces the FDPEF, which systematically deconstructs the multimodal metaphor competence of LLMs. Drawing inspiration from Gibbs' stages of metaphor processing, the FDPEF framework evaluates models across four key dimensions: understanding, generation, consistency, and creativity [13].

The FDPEF provides a comprehensive approach to evaluating the multimodal metaphor competence of large language models. By examining understanding, generation, consistency, and creativity, the framework offers a nuanced understanding of how models process and produce metaphors. This systematic evaluation is essential for advancing the capabilities of LLMs in both linguistic and multimodal contexts, contributing to the broader field of cognitive science and artificial intelligence.

3.3.1. Understanding Index (UI)

The understanding dimension is measured by semantic understanding accuracy (UIs), calculated as:

$$UI_s = \frac{1}{n}\sum_{i=1}^{n} \mathrm{II}\left(C_{ui} \in \Gamma(C_0)\right)$$

Where C_{ui} represents the i-th concept parsed by the system, $\Gamma(C_0)$ is the set of valid concepts, and $\mathrm{II}(\cdot)$ is an indicator function that takes 1 if the condition inside is met, otherwise 0. The capability mapping relationship for UI_s is divided into four levels:

Table 1. Capability mapping relationship for UI_s.

UI_s Range	Capability Level	Typical Performance
[0.9, 1]	Expert	Can recognize implicit concepts and puns in metaphors
[0.7, 0.9)	Proficient	Can parse literal metaphors accurately, partially understand cultural metaphors
[0.5, 0.7)	Basic	Can only handle conventional metaphors
[0, 0.5)	Deficient	Semantic misunderstanding or concept omission

Table 1 provides a detailed mapping of the Understanding Index score (UI_s) ranges to corresponding capability levels and typical performance characteristics in the context of multimodal metaphor comprehension.

3.3.2. Generation Index (GI)

The generation dimension is measured by modal coordination (GI_m), calculated according to the Bologna multimodal evaluation framework [21, 23]:

$$GI_m = \frac{1}{L}\sum_{l=1}^{L}(V_2T_l \times T_2V_l)$$

Where V_2T_l stands for the visual-to-textual alignment at layer l, and T_2V_l represents the textual-to-visual alignment at layer l, with values ranging from 0 to 1, defaulting to 3 layers, abstracted layer by layer. The capability mapping relationship for GI_m is divided into four levels, as shown in the Table 2 below:

Table 2. Capability mapping relationship for GI_m.

GI_m Range	Capability Level	Typical Performance
[0.8, 1]	Expert	Precisely describes details
[0.6, 0.8)	Proficient	Correct main subject and reasonable attributes
[0.4, 0.6)	Basic	Correct main subject but incorrect attributes
[0, 0.4)	Deficient	Severe semantic deviation

This table outlines the relationship between the Generation Index (GI_m) scores and the associated capability levels, along with typical performance in generating multimodal content.

3.3.3. Consistency Index (CI)

The consistency dimension considers the structural, semantic, and cognitive alignment between the original metaphor and the model-parsed regenerated metaphor. The overall consistency score is calculated as:

$$CI = \alpha \times CI_{struct} + \beta \times CI_{sem} + \gamma \times CI_{cog}$$

Where $\alpha + \beta + \gamma = 1$, with weights $\alpha = 0.4$, $\beta = 0.35$, and $\gamma = 0.25$. The rationale for this weight assignment, including both theoretical and empirical justifications, is detailed as follows.

Structural consistency (α=0.4, highest weight): CMT posits that the core of metaphor lies in the source-target domain mapping structure [18]. Without preserving this structural framework, the metaphor loses its fundamental validity—even if semantic or cognitive elements are partially aligned.

Semantic consistency (β=0.35, second-highest weight): Multimodal metaphor meaning is conveyed through the synergistic interaction of visual and textual semantics [11]. Semantic alignment ensures that the model retains the literal connotation and metaphorical implication of key elements (e.g., symbolic images, figurative language), which directly determines whether the regenerated metaphor conveys the original message.

Cognitive consistency (γ=0.25, lowest weight): Cognitive alignment reflects the match between the model's output and human metaphorical reasoning patterns [13]. While critical for naturalness, cognitive consistency is inherently more subjective and context-dependent than structural/semantic alignment—its relevance varies across cultural and individual cognitive differences—hence its lower theoretical priority relative to the other two dimensions.

Prior empirical studies on metaphor evaluation [3, 15] have consistently shown that structural mapping accuracy accounts for 35–45% of the variance in metaphor consistency scores, supporting the assignment of α=0.4 (within this empirical range).

Based on Lakoff & Johnson's conceptual mapping theory (1980), CI_{struct} is derived from the formulas:

$$CI_{struct} = \frac{1}{2}\left[\frac{I \times M}{|I| \cdot |M|} + \frac{T \cdot M\prime}{|T| \cdot |M\prime|}\right]$$

where I represents the generated image vector, T represents the generated text vector, M represents metaphor vector, M' represents the model-parsed regenerated metaphor vector. The capability mapping relationship for CI is shown in the table below:

Table 3. Capability mapping relationship for CI.

CI Range	Capability Level
[0.8, 1]	Human-level metaphor understanding
[0.7, 0.8)	Qualified
[0.6, 0.7)	Cognitive bias
[0, 0.6)	Metaphor mechanism failure

Table 3 details the Consistency Index (CI) scores, mapping them to capability levels in understanding metaphors.

3.3.4. Creativity Index (CrI)

The creativity dimension in language models is a crucial metric that assesses how novel or innovative a generated metaphor is compared to the original concept. This dimension is particularly important in evaluating the ability of models to generate content that is not only coherent but also creatively divergent from standard expressions. The creativity dimension is measured by conceptual span (CrI_d), calculated as:

$$CrI_d = \|E(C_0) - E(M)\|_2$$

In computational metaphor generation, conceptual span can be quantitatively measured by the Euclidean distance between semantic vectors, calculated as ‖E(C₀) - E(M)‖₂, where E(C₀) represents the semantic vector of the original concept and E(M) denotes that of the generated metaphor. A larger distance value reflects greater innovation in the generated metaphor, highlighting the model's ability to transcend conventional conceptual boundaries while preserving semantic coherence. This objective metric not only provides a quantitative basis for evaluating the creativity of language models but also informs strategies for the innovative integration of cross-modal information in multimodal generation tasks, thereby advancing model optimization and training for applications such as creative writing and advertising.

3.3.5. Scoring Procedure and Reliability

All scoring and quantitative evaluations in this study were conducted by a single independent annotator (the author) following a predefined, fixed rubric for each dimension (Understanding, Generation, Consistency, Creativity). The scoring rubric was established before evaluation and strictly followed the index formulas and level mapping rules presented in Sections 3.3.1–3.3.4.

To ensure transparency and replicability, all concepts, metaphor vectors, and alignment judgments were determined exclusively according to the computational formulas of UI, GI, CI, and CrI. Each score was derived from measurable indicators (semantic accuracy, modal alignment, structural mapping, conceptual span) rather than subjective intuition.

Since this study adopts a formula-driven quantitative evaluation rather than open subjective scoring, the procedural transparency and fixed decision rules ensure consistency and reproducibility. Future studies may invite multiple annotators and report inter-rater reliability to further validate the framework.

4. Results and Discussion

4.1. Performances of Large Language Models in Multimodal Metaphor Understanding

In the initial dialogue, the author used a public service advertisement released by the World Wildlife Fund, consisting of three images with hands painted to resemble a zebra, a crocodile, and a vulture, accompanied by the text "Give a hand to wildlife" in the upper left corner. Different LLMs provided varying responses: Claude-3-5 gave a more complete answer, while Cici's response was overly brief.

Overall, this advertisement effectively uses innovative visual design and a clear message to emphasize humanity's crucial role in wildlife protection. It encourages public action and engagement in conservation efforts, highlighting the interconnectedness of humans and the natural world."

After multiple similar dialogues, UI_s values for different LLMs were calculated.

Table 4. UI_s Values for different LLMs.

LLMs	UI_s
Cici	0.5
Kimi	0.7
GPT-4	0.8
Claude-3-5	0.9

Claude-3-5's analysis of the advertisement achieves a high UI_ s score of 0.9, indicating a sophisticated understanding of metaphorical content. The analysis effectively interprets the symbolic merger of human and animal features, emphasizing the interconnectedness between humans and nature. This aligns with the assertion that metaphors shape our understanding by linking disparate domains [18]. It captures the direct and engaging nature of the advertisement's message, encouraging public involvement. This reflects the persuasive function of metaphors which can powerfully influence attitudes and behaviors. It also appreciates the creative visual impact, recognizing how artistic elements draw attention and provoke reflection. This aligns with the exploration of metaphor in multimodal contexts, where visual elements enhance metaphorical meaning [11]. By highlighting the emotional appeal,

the analysis acknowledges the advertisement's ability to evoke personal responsibility, a concept supported by the emphasis on the role of emotion in metaphor comprehension [13]. It notes how familiar human elements combined with animal features create a memorable image, illustrating the cognitive blending that makes metaphors effective and thought-provoking [7]. Claude-3-5's analysis demonstrates a nuanced understanding of the advertisement's metaphorical elements, effectively integrating linguistic insights to provide a comprehensive interpretation.

UI_s results indicate that Claude-3-5 exhibits the highest proficiency in multimodal metaphor understanding, achieving an Understanding Index score (*UIs*) of 0.9. This score places Claude-3-5 at the expert level, demonstrating its superior ability to accurately recognize and interpret implicit concepts and double meanings within multimodal metaphors. On the other hand, Cici shows the weakest performance in this dimension, with a *UIs* of 0.5. This low score is attributed to the model's tendency towards over-abstract interpretations, which often leads to misunderstandings or omissions of key metaphorical concepts.

Traditional large language models (LLMs) suffer from the limitation of unimodal interaction (e.g., textual metaphor detection), while multimodal fusion is key to enhancing a model's general perceptual capabilities. The results of the index provide empirical support for this view. The success of Claude-3.5 demonstrates that when a language model possesses strong capabilities for integrating multimodal information, it can significantly improve performance on tasks such as metaphor understanding, which require cross-modal association and deep semantic comprehension. This directly validates the argument presented in the literature that "multimodal models can offer new possibilities for handling complex linguistic phenomena (e.g., metaphor)," indicating that multimodal technology is an effective path to overcome the shallow and abstract understanding of metaphors in LLMs. In contrast, Cici's failure cases reveal precisely the limitations warned of in the literature. If a model—even a multimodal one—cannot effectively coordinate information from different modalities, it may produce "over-abstract interpretations." This is a typical manifestation of the model's failure to successfully anchor and map concrete perceptual information (such as visual cues) to linguistic concepts, which is consistent with the limitation described as "lack of real-world perceptual grounding."

Metaphor processing capability is a key indicator for measuring deep semantic understanding and logical reasoning in models [22]. The UI evaluation results concretize and quantify this assertion. By translating the relatively abstract concept of "metaphor understanding ability" into a quantifiable "Understanding Index (UI)," the performance differences among various models become intuitive and comparable. Claude-3.5's "expert-level" performance indicates that it has reached a high level in semantic analysis and the recognition of implicit concepts and dual meanings. This aligns with the goal mentioned in the literature—that the GPT series achieves "deep metaphor modeling" through "multidimensional linguistic comparison"—but Claude-3.5 exhibits superior performance in a multimodal context. Cici's weak performance corroborates the "consistency" issue in current models' metaphor processing, as pointed out in the literature. The inconsistency in its outputs (sometimes correct, sometimes overly abstract) indicates instability in metaphor mapping and recognition, reflecting the immaturity of its internal cognitive architecture in handling metaphors.

The construction of multimodal metaphors relies on the interaction and integration between modalities and can be analyzed using theories such as visual grammar. The UI results verify the utility of these theoretical frameworks in evaluating a model's cognitive processes. Claude-3.5's high scores suggest that it may have intrinsically mastered a form of "modality coordination grammar," enabling it to effectively parse the interactive relationships between image elements and linguistic elements—much as described in visual grammar theory—thus accurately interpreting metaphors. Cici's "over-abstract interpretations" can be diagnosed as a failure in "conceptual integration." According to conceptual integration theory, the model may only operate within either the source or target domain, without constructing an effective blended space, thereby failing to derive the emergent meaning of the metaphor. This provides a theoretical perspective for diagnosing the reasons behind model failures.

The studies are actively exploring new methods to enhance models' metaphor capabilities [17, 30]. The UI results offer an important real-world reference and validation for these cutting-edge research efforts. The UI evaluation paradig m itself can be regarded as a form of benchmark testing similar to "CMT-based prompts." Its successful implementation demonstrates that constructing targeted evaluation frameworks is effective for measuring and promoting model progress. The results also point to future research directions: how to improve model architectures or training strategies (for instance, by drawing on Claude-3.5's successful experience) to avoid the "over-abstraction" issue exemplified by Cici. This provides insight for studies on collaborative visual metaphor generation by LLMs and diffusion models: the quality of generation is highly dependent on the quality of understanding, so core challenges at the comprehension level must be addressed first [4].

Except for Claude-3-5's score, GPT-4, UI_ s score of 0.8, ranks higher than other LLMs in understanding competence part, which aligns with the conclusion that "GPT-4 outperformed previous AI models on the Fig-QA dataset", although the competence is on novel literary metaphor interpretation [14].

4.2. Performances of Large Language Models in Multimodal Metaphor Generation

Compared to understanding capability, LLMs' generation capability is significantly weaker. The input instruction was to

generate an advertisement with the content "food safety commitment." Kimi's generated image was the most realistic, with high visual-to-text alignment but errors in textual-to-visual alignment, such as the word "commitment" not being displayed in the image. For the same request, Cici generated characters with obvious AI synthesis, unrelated to reality, while GPT-4 and Claude-3-5 generated content that mostly matched the text. The images generated by Cici and Kimi and their evaluations are as follows:

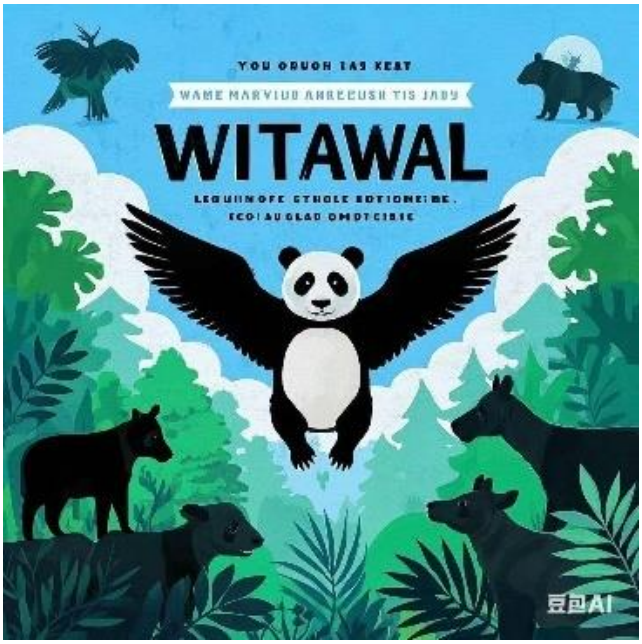

Figure 1. *Cici's image.*

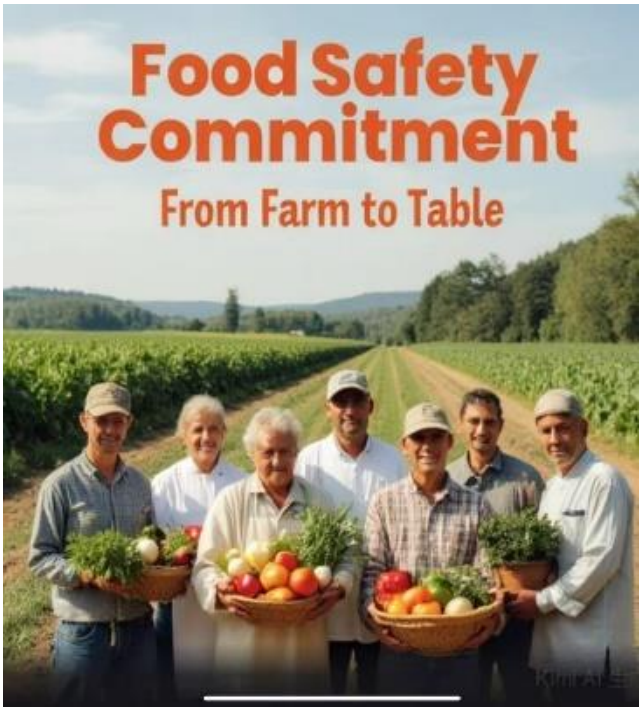

Figure 2. *Kimi's image.*

In comparison, GPT-4's matching was superior, exceeding 65% of food advertisement standards, with a complete logical chain: home scene → safety commitment → consumer decision, indicating more stringent procedural requirements.

After multiple commands for LLMs to generate different text and images and recording the results, GI_m values for different LLMs were calculated.

Table 5. *GI_m Values for different LLMs.*

LLMs	GI_m
Cici	0.007
Kimi	0.68
GPT-4	0.74
Claude-3-5	0.69

Cici's generated image effectively employs multimodal metaphor, showcasing distinct strengths and areas for improvement. The central panda symbolizes protection and harmony, creating a cohesive ecosystem with the surrounding animals and nature. This aligns with the idea that metaphors help us understand abstract concepts through concrete imagery [18]. The use of color and composition captures the viewer's attention, creating a serene and harmonious atmosphere. According to multimodal metaphors theory, visual elements significantly enhance information transmission [11]. The image evokes an emotional response, prompting viewers to reflect on environmental responsibility. Research emphasizes the role of emotion in enhancing the persuasive power of metaphors [13]. The abstract design of the text may hinder direct communication of the message. Existing literature highlights the need for a balance between linguistic and visual elements to ensure clarity and comprehensibility [16]. While the panda is a strong symbol of conservation in some cultures, it might not be as intuitive in others. This underscores the importance of cultural context in metaphor comprehension, as supported by conceptual blending theory [7].

This image generated by Kimi effectively uses multimodal metaphor to convey a message about food safety. The farmers holding fresh produce symbolize the direct connection and responsibility in ensuring food safety from the source. This aligns with the concept that metaphors help us understand complex ideas through tangible imagery [18]. The background of a farm leading into the distance emphasizes the journey of food from its origin to the table, reinforcing the "farm to table" concept. Visual metaphors can create powerful narratives that enhance understanding [11]. The diverse group of farmers underscores the universal importance of food safety, promoting inclusivity and shared responsibility. This reflects the idea of conceptual blending, where different elements merge to form a unified message [7]. While the text is clear, integrating it

more seamlessly with the visual elements could enhance cohesion. Existing research emphasizes the balance between linguistic and visual elements for more effective metaphorical communication [16]. The image could benefit from additional elements that evoke a stronger emotional connection [13].

The analysis of GPT-4's generated answers through mapping is shown in Figure 3.

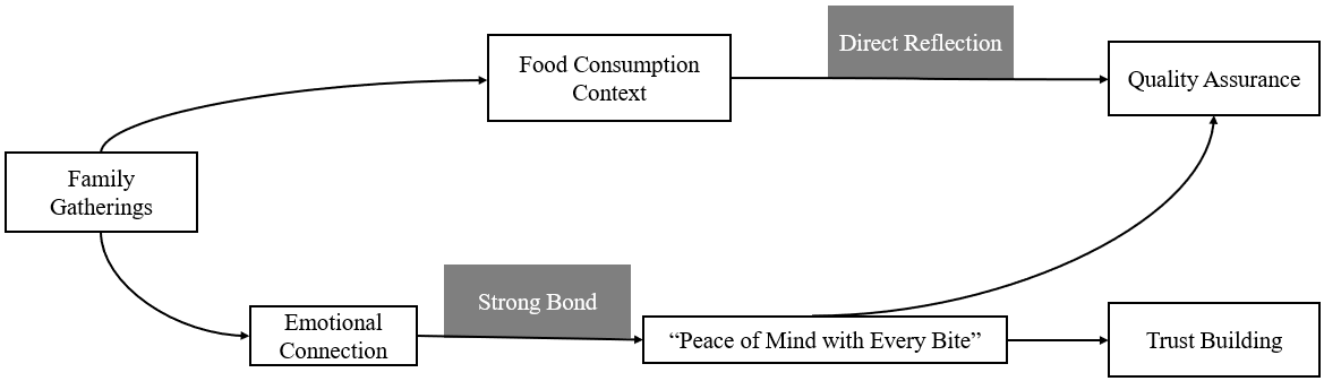

Figure 3. *Analysis of GPT-4's generated answers.*

The logical chain begins with "Family Gatherings," which sets a relatable and familiar context for consumers. This aligns with the assertion that metaphors structure our understanding of experiences by connecting abstract concepts with everyday life [18]. By establishing an "Emotional Connection," the content taps into the shared experiences of family meals, creating a "Strong Bond" with the audience. Research emphasizes that effective metaphors often evoke emotional responses, enhancing their persuasive power [13]. The transition from "Food Consumption Context" to "Quality Assurance" and "Trust Building" reflects a direct reflection of the brand's commitment to safety. Visual and textual metaphors can reinforce trust by clearly communicating reliability and quality [11]. The phrase "Peace of Mind with Every Bite" encapsulates the assurance provided by the brand, influencing consumer decisions by highlighting safety and trust. The concept of conceptual blending illustrates how combining familiar elements can lead to new insights and reinforce consumer confidence [7].

In terms of generation capability, GPT-4 outperforms other models with a Generation Index (*GIm*) of 0.74, which corresponds to a proficient level. This indicates that GPT-4 excels in multimodal mapping logic, effectively coordinating visual and textual information to generate coherent and contextually appropriate metaphors. However, it is important to note that despite this proficiency, all evaluated models, including GPT-4, struggle with the "semantic deviation" problem. This issue arises when there is a misalignment between the generated text and image, leading to inconsistencies that detract from the overall coherence of the multimodal output.

4.3. Performances of Large Language Models in Multimodal Metaphor Consistency

First, GPT-4 was used for consistency dimension detection. The first step involved parsing an IKEA quilt advertisement (Figure 4), recording the first response. The second step used the parsed content to generate an advertisement (Figure 5), then had GPT-4 read it, recording the second response.

Figure 4. *IKEA's advertisement.*

Figure 5. *Image generated by GPT-4.*

After consistency analysis, the overall consistency index was calculated as 0.703. And GPT-4's report is as follows:

Table 6. *GPT-4's Consistency Analysis.*

Evaluation Dimension	Ad A (Energy Pill)	Ad B (Rechargeable Battery)	Match Degree
Structural alignment	Consumer symbols	Technology symbols	0.65
Semantic alignment	Energy conversion theory	Physiological restoration theory	0.78
Cognitive alignment	Life solution provider	Technology enabler	0.61

Similarly, other LLMs were trained with corresponding dialogues, recording generated metaphor vectors and calculating CI to draw conclusions (Table 6). The generated text and images from the second iteration of LLMs were used for creativity quantification. For example, GPT-4 generated a battery charging advertisement that ranked in the top 25% of similar technological metaphors, with the source domain being a good night's sleep for humans and the target domain being battery charging. The analysis of GPT-4's metaphorical constructs, such as the battery charging advertisement, highlights the model's ability to leverage familiar human experiences to convey complex technological concepts. This aligns with the theory that metaphors allow us to understand one domain of experience in terms of another, facilitating comprehension and relatability [18]. Despite achieving a creativity score of 0.607, indicating a high level of innovation within predefined parameters, the creativity of LLMs remains constrained by their reliance on existing human cognitive frameworks. The theory of conceptual blending suggests that while LLMs can combine different domains to generate novel ideas, they are inherently limited by the inputs and patterns they have been trained on [7]. The path dependency observed in GPT-4's metaphorical constructs, where charging corresponds to energy replenishment akin to sleep, underscores the model's reliance on established cognitive pathways. This reflects the notion of combinatorial creativity, where new ideas are generated through the recombination of familiar concepts rather than the creation of entirely new paradigms [2].

When assessing consistency, GPT-4 achieves a Consistency Index (CI) of 0.703. This score suggests that GPT-4 is approaching the human-level threshold for metaphor comprehension, indicating a high degree of structural, semantic, and cognitive alignment between the original metaphor and

the model's parsed and regenerated metaphors. However, the presence of cognitive biases in GPT-4's outputs highlights areas where the model's understanding and generation processes may still diverge from human cognitive patterns.

4.4. Performances of Large Language Models in Multimodal Metaphor Creativity

As demonstrated in Table 7, while LLMs like GPT-4 can simulate creative processes, true breakthroughs in creativity may require advancements in how these models conceptualize and synthesize information beyond existing human cognition. As pointed out, achieving transformational creativity—where new conceptual spaces are explored—remains a challenge for artificial intelligence [2].

Table 7. CI and CrI_d Values for different LLMs.

LLMs	CI	CrI_d
Cici	0.413	0.352
Kimi	0.675	0.497
GPT-4	0.703	0.607
Claude-3-5	0.654	0.483

In the dimension of creativity, the evaluated LLMs generally demonstrate a reliance on conventional metaphor paradigms. For instance, GPT-4 achieves a Creativity Index (*CrId*) of 0.607. While this score reflects a moderate level of innovation, it also underscores the limitations imposed by the inherent cognitive frameworks embedded within the training data. These limitations restrict the models' ability to generate novel and unconventional metaphors, often resulting in outputs that adhere to familiar and predictable metaphorical constructs.

5. Conclusion

5.1. Major Findings

This study examines the performance of several large language models on multimodal metaphor processing using the proposed Four-Dimension Progressive Evaluation Framework (FDPEF), which assesses understanding, generation, consistency, and creativity. The key findings are summarized as follows.

First, Claude-3.5 outperformed all models in multimodal metaphor understanding, with an Understanding Index (UIs) of 0.9, reaching an expert-level interpretation of implicit concepts, symbolic structures, and emotional connotations in multimodal metaphors. By contrast, Cici achieved the lowest score (UIs = 0.5), as its frequent over-abstraction resulted in inaccurate or incomplete interpretation of core metaphorical meanings. These results highlight the importance of balanced abstraction and precision in metaphor comprehension.

Second, GPT-4 exhibited the strongest metaphor generation capability with a Generation Index (GIm) of 0.74, showing coherent multimodal mapping and effective coordination of visual and textual information. Nevertheless, all models suffered from semantic deviation, where mismatches between generated text and images impaired overall coherence. Although Cici produced visually and culturally salient metaphors such as the panda for environmental protection, it struggled to balance visual and linguistic clarity. Cultural variability also affected metaphor interpretability across audiences.

Third, in terms of consistency, GPT-4 achieved a CI score of 0.703, approaching human-level performance in maintaining structural and semantic alignment between original, parsed, and regenerated metaphors. Even so, subtle cognitive biases still caused occasional divergences from human-like reasoning patterns.

Fourth, regarding creativity, all evaluated LLMs remained constrained by conventional metaphor schemas. GPT-4 obtained a CrId of 0.607, indicating only moderate innovation, as its outputs typically combined familiar concepts rather than generating truly novel mappings. Similar limitations were observed in other models, whose metaphor generation was largely restricted by path dependence and the cognitive frameworks embedded in training data.

5.2. Implications of the Study

The implications of this research extend well beyond technical artificial intelligence and reach into a series of interdisciplinary fields, including cognitive linguistics, multimodal discourse analysis, computational communication, and intelligent creative design.

In cognitive science, investigating how large language models process, understand, and generate multimodal metaphors provides a new computational perspective for re-examining human conceptual systems and metaphorical reasoning mechanisms. By comparing the similarities and differences between LLM-based metaphor processing and human cognitive patterns, this study helps refine existing theories of conceptual mapping, cross-domain association, and conceptual blending. In particular, the observed patterns of consistency and deviation in model performance can inspire more dynamic, context-sensitive models of human metaphor comprehension, moving beyond static conceptual metaphor frameworks toward accounts that better reflect real-world situated cognition.

For the field of multimodal discourse analysis, this study offers a quantifiable, data-driven approach to examining metaphor construction across visual and linguistic modalities. Traditional multimodal metaphor research often relies on qualitative interpretation and manual annotation, which introduces subjectivity and limits comparability. The Four-Dimensional Progressive Evaluation Framework (FDPEF) proposed in this study

provides a systematic, replicable method for evaluating metaphor understanding, generation, consistency, and creativity. This framework can be adopted as a methodological reference for future empirical studies, helping to bridge qualitative linguistic analysis and quantitative computational evaluation.

In practical terms, the findings carry meaningful implications for applied fields such as advertising design, public communication, cross-cultural communication, and human-computer interaction. For advertising and brand communication, understanding the strengths and limitations of LLMs in multimodal metaphor generation can support the development of more persuasive, emotionally resonant, and culturally appropriate creative content. For cross-cultural communication, the observed sensitivity of metaphors to cultural backgrounds highlights the necessity of incorporating cultural knowledge into multimodal models, ensuring that metaphorical expressions are not only linguistically coherent but also culturally acceptable and contextually effective. In human-computer interaction systems, improved metaphor processing can enhance the naturalness, interpretability, and emotional engagement of intelligent interfaces, making machine communication more human-like and intuitive.

Despite these promising implications, several challenges must be addressed to fully unlock the potential of LLMs in multimodal metaphor processing.

First, the effective integration of heterogeneous multimodal data remains a fundamental challenge. Visual features, textual semantics, emotional tones, and contextual cues must be fused at a deep cognitive level rather than merely combined at a surface level, requiring more advanced cross-modal alignment architectures and joint representation learning mechanisms.

Second, cultural variability in metaphor understanding and usage presents a persistent barrier [16]. Metaphors that are intuitive and persuasive in one cultural context may be ambiguous, weak, or even counterproductive in another. Future multimodal models should be equipped with adaptive cultural understanding modules to adjust metaphorical mappings according to specific cultural schemas and value systems.

Third, the trade-off between conventionality and creativity in metaphor generation requires further exploration. Current LLMs tend to rely on established, conventional metaphor patterns due to training data constraints, which limits their ability to produce truly innovative and transformative metaphors. To achieve higher creativity, future models need to integrate dynamic cognitive blending mechanisms and break through path dependence in conceptual association.

Addressing these challenges will not only improve the performance of LLMs in figurative language processing but also deepen the theoretical dialogue between artificial intelligence, cognitive linguistics, and multimodal communication.

5.3. Limitations of the Study

The study's approach to evaluating the creativity and effectiveness of LLM-generated metaphors presents several limitations, primarily due to the inherent subjectivity in scoring standards and the current capabilities of AI models. These limitations highlight the challenges and potential areas for future research and development.

Although semantic vectors were quantified using Transformer models, the division of semantic hierarchies required manual analysis. This introduces subjectivity, as evaluators' interpretations of metaphorical content can vary significantly. Metaphors are deeply influenced by cultural and individual cognition, which can lead to varied understandings and applications [18]. The subjective nature of metaphor interpretation is further complicated by cultural differences. Metaphors that resonate in one cultural context may not hold the same meaning in another. Cultural variations can lead to different metaphorical expressions and understandings, impacting the consistency of metaphor evaluation across diverse contexts [16]. The reliance on manual analysis in dividing semantic hierarchies can lead to score deviations. This variability underscores the need for more objective and standardized evaluation methods to ensure consistency and reliability in metaphor assessment. The LLMs studied, such as GPT-4 and Claude-3-5, excel in generating and understanding linguistic metaphors. However, they do not represent the full spectrum of generative AI models. This limitation suggests that current models may not fully capture the complexity and nuance of multimodal metaphors, which involve both visual and textual elements. Existing research indicates that multimodal metaphor understanding and generation require advanced models capable of integrating diverse types of information. Multimodal models show clear advantages in understanding complex metaphors, suggesting that future LLMs need to incorporate visual data to improve metaphor processing [15, 32]. Current LLMs are constrained by the cognitive frameworks upon which they are based. Conceptual blending relies on existing mental spaces, limiting the depth and novelty of generated metaphors [7]. This constraint suggests that LLMs may struggle to produce truly innovative metaphors without significant advancements in cognitive modeling.

5.4. Suggestions for Further Research

To address the subjectivity in scoring, future research should focus on developing more objective evaluation methods. This could involve the use of automated systems capable of analyzing metaphorical content without human bias, ensuring more consistent and reliable results. Advancements in Multimodal Models: The development of advanced multimodal LLMs is crucial for improving metaphor understanding and generation. By integrating visual and textual information, these models can better capture the complexity of metaphors, enhancing both accuracy and creativity. Multimodal models offer significant advantages in understanding and generating complex metaphors [15].

Future models should also incorporate mechanisms for cultural sensitivity, allowing them to adapt metaphor generation and understanding to different cultural contexts. This adaptation would enhance the relevance and applicability of metaphors across diverse audiences, addressing the cultural limitations [16]. To achieve breakthroughs in creativity, LLMs must transcend current cognitive limitations and develop mechanisms for generating novel conceptual spaces. This could involve integrating more dynamic learning models that mimic human creative processes more closely, as noted in discussions on creativity [2].

In conclusion, the study underscores the significant progress made by LLMs in the realm of metaphor understanding and generation, while also highlighting the potential for future advancements. By embracing the possibilities offered by multimodal models and addressing the inherent challenges, researchers can unlock new avenues for innovation in artificial intelligence and cognitive science. As we continue to explore the intricate interplay between language and imagery, the potential for LLMs to transform our understanding of metaphors remains vast and exciting.

Abbreviations

LLMs	Large Language Models
FDPEF	Four-Dimension Progressive Evaluation Framework
MDA	Multimodal Discourse Analysis
CMT	Conceptual Metaphor Theory

Author Contributions

Zhong Yuke: Conceptualization, Data curation, Formal Analysis, Funding acquisition, Investigation, Methodology, Project administration, Resources, Software, Supervision, Validation, Visualization, Writing – original draft, Writing – review & editing

Conflicts of Interest

The author declares no conflicts of interest.

References

[1] Agerri, R., & Sanchez-Bayona, E. (2025). Metaphor and large language models: When surface features matter more than deep understanding. *Findings of the Association for Computational Linguistics: ACL 2025*, 17462–17477. https://doi.org/10.48550/arXiv.2507.15357

[2] Boden, M. A. (2004). *The creative mind: Myths and mechanisms*. Routledge. https://doi.org/10.4324/9780203508527

[3] Charteris-Black, J. (2004). *Corpus approaches to critical metaphor analysis*. Palgrave Macmillan.

[4] Chakrabarty, T., et al. (2023). I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *Findings of the Association for Computational Linguistics: ACL 2023*, 7370-7388, Toronto, Canada. Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.findings-acl.465

[5] Chen, H. M., Liu, Z. Y., & Sun, M. S. (2024). The social opportunities and challenges in the era of large language models. *Journal of Computer Research and Development*, 61(05), 1094-1103.

[6] Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv: 1810.04805*. https://doi.org/10.48550/arXiv.1810.04805

[7] Fauconnier, G., & Turner, M. (2002). *The way we think: Conceptual blending and the mind's hidden complexities*. Basic Books. https://doi.org/10.5860/choice.40-1223

[8] Feng, D. Z., Zhang, D. L., & Kay O'Halloran. (2014). Advances and frontiers of multimodal discourse analysis. *Contemporary Linguistics*, 16(01), 88-99+126.

[9] Feng, Z. W., & Ding, X. M. (2024). Development of AI and alignment of LLMs. *Journal of Language Governance*, 1(01), 108-126.

[10] Feng, Z. W., & Zhang, D. K. (2023). GPT and language research. *Technology Enhanced Foreign Language*, 40(02), 3-11+105.

[11] Forceville, C. (1996). *Pictorial metaphor in advertising*. Routledge. https://doi.org/10.4324/9780203064252

[12] Forceville, C., & Urios-Aparisi, E. (2009). *Multimodal metaphor*. De Gruyter. https://doi.org/10.1515/9783110215366

[13] Gibbs, R. W. (2006). *Embodiment and cognitive science*. Cambridge University Press. https://doi.org/10.1017/CBO9780511805844

[14] Ichien, N., Stamenković, D., & Holyoak, K. J. (2024). Large language model displays emergent ability to interpret novel literary metaphors. *Metaphor and Symbol*, 39(4), 296-309. https://doi.org/10.48550/arXiv.2308.01497

[15] Kiros, R., Salakhutdinov, R., & Zemel, R. (2015). Unifying visual-semantic embeddings with multimodal neural language models. In *Advances in Neural Information Processing Systems (NeurIPS)*. https://doi.org/10.48550/arXiv.1411.2539

[16] Kövecses, Z. (2005). *Metaphor in culture: Universality and variation*. Cambridge University Press. https://doi.org/10.1017/CBO9780511614408

[17] Kramer, O. (2025). Conceptual metaphor theory as a prompting paradigm for large language models. *Computer and Language: arXiv: 2502.01901*.

[18] Lakoff, G., & Johnson, M. (1980). *Metaphors we live by*. University of Chicago Press.

[19] Li, G., Wang, Z. S., He, X. T., et al. (2023). From ChatGPT to large multimodal model: Present and future. *Bulletin of National Natural Science Foundation of China*, 37(05), 724-734.

[20] Liu, M., Wu, Z. M., Liao, J., et al. (2023). Educational applications of large language models: Principles, status and challenges—from light-weighted BERT to conversational ChatGPT. *Modern Educational Technology*, 33(08), 19-28.

[21] Pelosi, G., et al. (2023). Bologna multimodal evaluation framework.

[22] Qin, H. W., & Zhou, X. (2024). Comparative research on large language models and languages. *Foreign Language Teaching and Research*, 56(02), 163-176+318.

[23] Sun, Y. (2012). Metaphorical research based on the corpus tool Wmatrix. *Foreign Language Education*, (03), 7-11.

[24] Teng, D., & Miao, X. W. (2018). Meaning construction of multi-modal metaphors in the picture book discourse from the grammar of visual design. *Foreign Language Research*, (05), 53-59.

[25] Wang, L. H., & Liu, X. Y. (2013). Interpretation of metaphorical multimodal discourse based on conceptual blending theory. *Technology Enhanced Foreign Language Education*, (06), 28-33.

[26] Wu, A. P., & Zhong, S. M. (2014). Multimodal discourse research on visual grammar and metaphorical mechanism. *Foreign Languages and Their Teaching*, (03), 23-28.

[27] Xie, J. X. (2011). A multimodal perspective on metaphor—Based on a review of metaphor studies of Charles Forceville. *Foreign Language Research*, (05), 49-52.

[28] Xu, Y. M., Hu, L., Zhao, J. Y., et al. (2024). Technology application prospects and risk challenges of large language models. *Journal of Computer Applications*, 44(06), 1655-1662.

[29] Yang, Y. W. (2015). Poster discourse multimodal metaphorical representation type research. *Foreign Languages Research*, (03), 30-35.

[30] Zhang, D.-Y., et al. (2025). Towards multimodal metaphor understanding: A Chinese dataset and model for metaphor mapping identification. *ACM Transactions on Asian and Low-Resource Language Information Processing*. Advance online publication.

[31] Zhang, Z., et al. (2025). Cross-modal information flow in multimodal large language models. *Artificial Intelligence: arXiv: 2411.18620*.

[32] Zhao, C. Y., Zhu, G. B., & Wang, J. Q. (2023). The inspiration brought by ChatGPT to LLM and new development ideas of multi-modal large model. *Data Analysis and Knowledge Discovery*, 7(03), 26-35.

[33] Zhao, X. F. (2011). New developments in conceptual metaphor research: Multimodal metaphor research: A review of Forceville & Urios-Aparisi multimodal metaphor. *Foreign Languages Research*, (01), 1-10+112.