Scientific Research Publishing

Open Journal of Social Sciences, 2025, 13(8), 474-488
https://www.scirp.org/journal/jss
ISSN Online: 2327-5960
ISSN Print: 2327-5952

# Multimodal Discursive Construction of Inner Mongolian Image: Taking the Documentary *Charming Inner Mongolia* as an Example

**Hecong Wang, Haicui Zheng**

Foreign Languages College, Inner Mongolia University, Hohhot, China
Email: 1104059279@qq.com

**How to cite this paper:** Wang, H. C., & Zheng, H. C. (2025). Multimodal Discursive Construction of Inner Mongolian Image: Taking the Documentary *Charming Inner Mongolia* as an Example. *Open Journal of Social Sciences, 13,* 474-488.
https://doi.org/10.4236/jss.2025.138031

**Received:** July 28, 2025
**Accepted:** August 17, 2025
**Published:** August 20, 2025



## Abstract

This study employs Halliday's Systemic Functional Grammar and Kress & van Leeuwen's Visual Grammar as theoretical foundations, integrated with Zhang Delu's multimodal discourse analysis framework, to investigate the construction of Inner Mongolia's ecological image in the documentary *Charming Inner Mongolia*. The research analyzes multimodal strategies across Ideational, Interpersonal, Textual function and Representational, Interactive, Compositional meanings with inter-semiotic relations among multimodal resources. The documentary portrays Inner Mongolia through six defining ecological attributes: primordial ecosystem guardian, nomadic wisdom transmitter, natural sanctity venerator, ethnic aesthetics interpreter, green ecology perserverer, and sustainable development practitioner. Findings demonstrate that the strategic complementarity of multimodal symbols (linguistic, visual, and auditory) effectively reinforces image.

## Keywords

Multimodal Discourse Analysis, Visual Grammar, The Documentary *Charming Inner Mongolia*, Image Construction

## 1. Introduction

With the deepening implementation of the 14th Five-Year Plan (2021-2025) and the gradual advancement of the 2035 long-range objectives, constructing and disseminating Inner Mongolia's regional image serves as both a vital window showcasing China's multicultural diversity and a key drive for regional economic growth and cultural exchange. Recent years have witnessed multimodal discourse analysis theory offering innovative perspectives for interpreting regional representations.

As a distinctive ethnic region abundant in cultural resources, Inner Mongolia presents a signature cultural asset to national and global audiences (Li, 2019). Regional image falls within the scope of brand identity (Jiang, Zhu, & Lu, 2006). Despite Inner Mongolia's rich natural endowments and unique ethnic heritage, academic research remains limited, particularly regarding image construction through multimodal discourse, which lacks systematic theoretical frameworks and empirical analysis.

The documentary *Charming Inner Mongolia* utilizes multimodal resources including imagery, sound, and text to showcase the region's natural landscapes, historical culture, and modern development, providing rich textual material for research. Grounded in Halliday's Systemic Functional Grammar and Kress & van Leeuwen's Visual Grammar, and integrated with Zhang Delu's multimodal discourse analysis framework, this study examines the construction of Inner Mongolia's ecological image in the documentary. The research aims to provide theoretical references and practical guidance for enhancing the dissemination and promotion of Inner Mongolia's regional identity.

## 2. *Charming Inner Mongolia*: A Typical Case of Multimodal Discourse

The documentary *Charming Inner Mongolia* was jointly produced by China.org.cn and the Internet Information Office of Inner Mongolia Autonomous Region one year after the 70$^{th}$ anniversary of Inner Mongolia. It was released on August 9, 2018, and immediately drew enthusiastic responses and high praise. The entire documentary lasts for 43 minutes, mainly in English with Chinese subtitles, comprehensively showcasing Inner Mongolia's folk culture, ecological landscapes, and the people's morale and cultural vitality. The documentary is divided into eight episodes, each with a corresponding subtitle: Horse-head Fiddle Tunes with the New World, the Sound of History: The Ancient Mongolian Song, Youth Football in Inner Mongolia: Play with Passion, Jirgaltu: I Won the 40 Years' Battle against the Desert, Wei Gang: Public Service Volunteering Needs No Word of Promises, Mahai Master: Composing Songs of Life with Needle and Thread, Ulan Mugir Troupe 60 Years On, and the Three Manly Skills of Hero Bateer.

Visually, the documentary features a large number of scenes with distinct Inner Mongolian characteristics, such as the changing seasons of the grassland, dynamic scenes of the Nadam Fair, and contrast shots between yurts and modern cities. It uses various shooting techniques like aerial shots and close-ups to enhance the regional identity. In terms of language modality, English is the main narration language, supplemented by bilingual subtitles in Chinese and English, which not only expands the acceptance range of international audiences but also avoids meaning loss in translation, providing a more accurate basis for multimodal discourse analysis. In the auditory modality, the narration and commentary adopt a calm and objective tone, while the background music extensively uses Inner Mongolian-specific instruments like the horsehead fiddle and singing forms like throat

singing, strengthening the ethnic uniqueness. In the documentary, the interaction among the three modes is frequent. For instance, the visual mode is often combined with the language mode, such as the galloping horses accompanied by the caption "Horses are intimate pals, family members, and battle companions" to showcase the nomadic wisdom of Inner Mongolia. The scene of people gathering together for a meal, combined with the background music, is used to create an atmosphere of transitioning from sadness to joy. Or the explanatory voiceover and the textual modality in the auditory aspect provide the audience with complete information, etc. In summary, the documentary *Charming Inner Mongolia* presents an exemplary case for multimodal discourse analysis in constructing regional ecological imagery.

## 3. Theoretical Framework

Multimodal discourse refers to communicative phenomena that engage multiple sensory channels (auditory, visual, tactile, etc.) through diverse semiotic resources such as language, images, sound, and gesture (Zhang & Wang, 2011). In 1994, Halliday (1994) posited language as a social semiotic process, asserting that all social semiotics generate and transmit meaning. Systemic Functional Grammar (SFG), developed by Halliday, is a significant framework in linguistics. It includes systemic and functional grammar, where systemic grammar explains language's internal structure as a network of meaning potentials, and functional grammar focuses on language's natural functions (Halliday & Mattehiessen, 2004). According to SFG, language simultaneously realizes three metafunctions: Ideational, Interpersonal, and Textual metafunctions (Halliday, 1973). The ideational function describes how language represents the objective world, including participants, processes, and circumstances. Halliday identified six processes within the transitivity system: material, relational, mental, verbal, behavioral, and existential processes. The interpersonal function deals with how language expresses social and personal relationships, achieved primarily through mood and modality. Mood indicates the speaker's intent, whether asking questions or making statements, while modality conveys the speaker's attitude and the likelihood or obligation of the proposition. The textual function involves forming coherent texts through various text formation means, categorized into structural and non-structural cohesion. It includes theme, information point, and rheme, with marked and unmarked themes indicating topic overlap or inconsistency.

Inspired by Halliday, Kress and van Leeuwen proposed three corresponding meanings for visual analysis: Representational, Interactive, and Compositional meanings—called Visual Grammar (Kress & Van Leeuwen, 2021). Representational meaning divides into narrative and conceptual representations—the former features dynamic temporal sequences while the latter depicts static classifications. Diagonal alignments of visual elements form vectors, which serve as the fundamental construct of narrative representation (Van Ieeuwen, 2005). Narrative representation conveys dynamic processes through three action types: reactional pro-

cesses (gaze vectors), verbal/mental processes (thought/speech bubbles), and actional processes (bodily movement vectors) (Li, 2003). Conversely, conceptual representation presents participants in generalized or stable states without vectors or temporal sequences. "It is a generalized, stable, and timeless essence" (Wei, 2008). Interactive meaning addresses "the interpersonal relationship between represented participants and viewers, realized through contact, modality, social distance, and attitude" (Hu, 2017). Compositional meaning integrates representational and interactive dimensions by examining how salience, information value, and framing organize visual semiosis into coherent wholes (Tian & Zhang, 2013).

In 2009, Zhang Delu established a comprehensive multimodal discourse analysis framework based on SFG. This framework comprises four levels: the cultural level, the contextual level, the content level, and the expression level. In content level, it can be divided into Discourse meaning and Forms and Relations. Zhang Delu categorizes intermodal relations into complementary and non-complementary types, with complementary relations further subdivided into reinforced relations and unreinforced relations (Zhang, 2009). Complementarity in multimodal relations explores how different modalities interact to construct and reinforce meaning. In the reinforced relationship, one mode is the main form of communication, and the other mode(s) reinforce it. In an unreinforced relationship, both communication modes are indispensable and complementary to each other. The theoretical framework adopted is visualized in **Figure 1**.

This study establishes a three-dimensional analytical framework based on Kress and Van Leeuwen's Visual Grammar, Halliday's Systemic Functional Grammar,

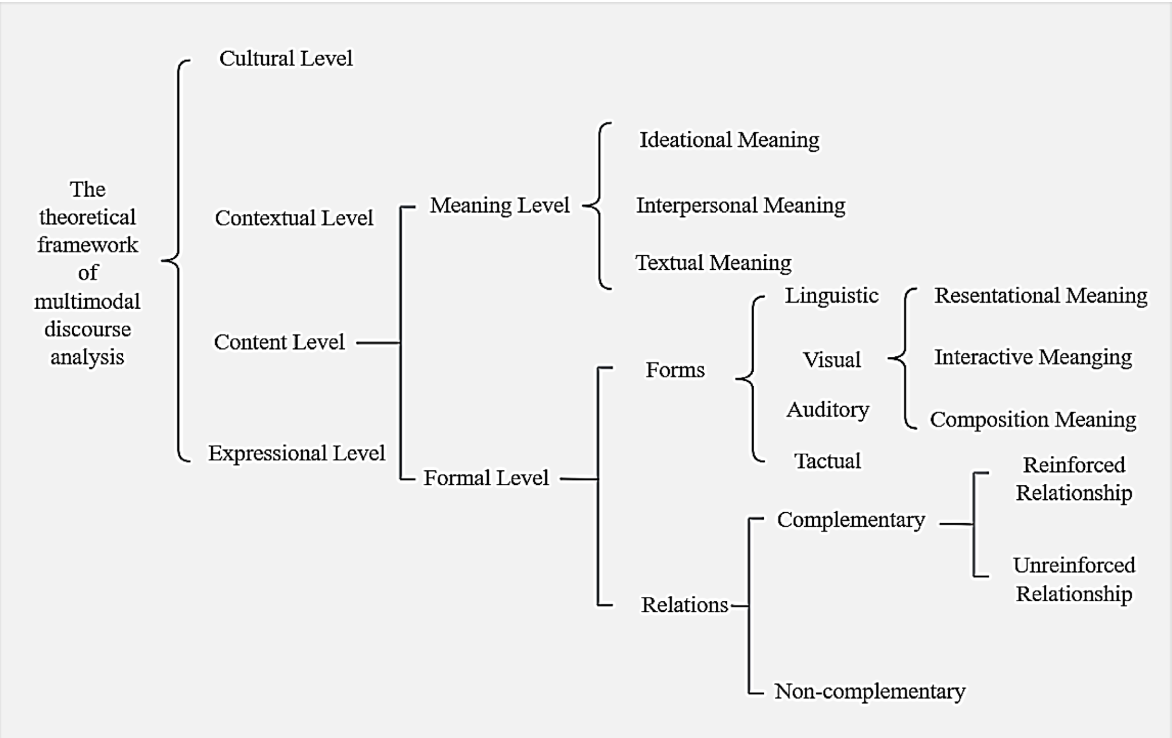


**Figure 1.** Multimodal discourse analysis theoretical framework.

and Zhang Delu's theory of multimodal complementarity to examine the multimodal construction mechanisms of Inner Mongolia's ecological image in the documentary.

First, Systemic Functional Grammar deciphers linguistic modalities through ideational function, interpersonal function, and textual function. Second, Visual Grammar provides analytical tools across three dimensions: representational meaning, interactive meaning, and compositional meaning. Finally, Zhang Delu's theory of multimodal complementarity examines cross-modal synergy: reinforcement relations and non-reinforcement relations. The integrated analytical model has been deliberately selected because it offers a comprehensive, hierarchical, and dynamic toolkit uniquely suited to the documentary's ecological-communication aims. SFG explicates how verbal resources construe experience, enact social relations, and weave coherent texts; Visual Grammar translates these semiotic principles into the visual domain, capturing how images narrate, address viewers; Zhang Delu's theory then distinguishes between reinforcement and non-reinforcement complementarity among modes. This synthesis allows the study to move seamlessly from micro-level grammatical choices to macro-level ideological positioning, ensuring that the ecological image of Inner Mongolia is analysed both in its monomodal specificities and in the synergistic orchestration that ultimately shapes audience perception.

## 4. The Multimodal Discourse Construction of Inner Mongolia's Ecological Image in the Documentary

### 4.1. Guardian of Primordial Ecosystems

Inner Mongolia, as the autonomous region spanning the widest range of latitudes in China, features a diverse range of natural landscapes that vary progressively with latitude. These include dense forests, expansive grasslands, and vast desert areas.

In the second episode of the documentary, a shooting technique that moves from east to west is employed, mirroring the actual geomorphic distribution and showcasing the natural landscapes of Inner Mongolia. In this context, the green forests, grasslands, tawny deserts, and blue sky are contrasted in tone. The camera continuously displays three images, constantly contrasting the dark green primeval forests, the green waves of the tender grasslands, and the large color blocks of the yellow desert wonders, all showing a colorful image of Inner Mongolia. The modality is used to assess the authenticity and credibility of the image, including high modality, medium modality, and low modality (Li, 2003). The color of the image in high modality is natural and has high saturation. The color saturation of the image in medium modality is insufficient. The color of the image in low modality is black and white (Van Ieeuwen, 2005). **Figures 2-4** are presented in very bright colors to showcase the natural scenery of Inner Mongolia. They reflect high modality within the interactive meaning of visual grammar. This approach creates an impression of vibrancy and vitality for the audience.


**Figure 2.** The great forest.


**Figure 3.** The great grassland.


**Figure 4.** The great desert.

At this time, the subtitle states, "The biggest feature here is that it has a great forest, a great grassland, and a great desert." The interactive function in the Systemic Functional Grammar involves the role of both the speaker and the listener. It emphasizes the "great forest", "great grassland", and "great desert" of Inner Mongolia. By repeatedly highlighting these three "great" features through absolute expressions and emotional cues, it praises the natural scenery of Inner Mongolia, directs the audience's attention to the region's natural characteristics, and encourages them to accept the emotional implications. This helps construct a beautiful image of Inner Mongolia's natural environment. From the perspective of the content level in Zhang Delu's framework, different modalities exhibit distinct relationships at the formal level. Since both modalities are indispensable for

the construction of overall meaning, they form a complementary relationship. The two modalities form an auxiliary reinforcement relationship in a complementary relationship. With the visual modality as primary, employing vibrant colors to showcase Inner Mongolia's pristine natural landscapes, and the textual modality as secondary to supplement the imagery, they synergistically present the region's unadorned beauty. This intermodal collaboration constructs the ecological image of Inner Mongolia as a primordial guardian of wilderness.

### 4.2. Transmitter of Nomadic Ecological Wisdom

During the thousand-year-long process of production, the Mongolian people have developed a set of their own survival philosophy. When the grassland style is displayed (**Figure 3**), the white flock of sheep is placed at the center of the picture, which creates strong salience against the green grassland background. Salience indicates that the components in an image attract the viewer's attention to varying degrees, which can be achieved through placement in the foreground or background, relative size, contrast of tone values, and differences in sharpness (Yuan & Nai, 2022). With a high degree of salience, an element easily attracts the viewer's attention; conversely, with a low degree of salience, it is difficult to draw attention. Placing the sheep in a prominent central position emphasizes the important role of animals in the ecology of Inner Mongolia, and also illustrates the original ecological nomadic culture, deepening the audience's impression of the Inner Mongolia grasslands. In addition, **Figure 3** shows us the traditional nomadic grazing in Inner Mongolia throughout history. The mobility of the herds not only prevented excessive grazing but also utilized animal manure and trampling for natural fertilization, achieving the regeneration of the grassland. This is an ecological wisdom of the people of Inner Mongolia.

The narrative representation of the action process primarily involves three elements: the Actor, the Goal, and the Vector. The Actor is typically the main participant initiating the Vector, which constitutes the specific action performed by the Actor or the Actor itself. The participant towards whom the Vector is ultimately directed is the Goal. The galloping horses constitute an action process (**Figure 5**). Here, the horse serves as the Actor, the upward lift of its hooves functions as the Vector, and with no specific Goal identified, this constitutes an intransitive action process. In the scene, the horse bears no saddle, running unbridled across undulating waves of grass. This vividly captures the natural rhythms unique to the grasslands and the untamed beauty free from human discipline. From an ideational function perspective, the subtitle "horses are intimate pals, family members, and battle companions" elevates the horse from mere livestock to a core member within the grassland ecological network. The agility and loyalty of horses are demonstrated through three subordinate clauses. During the migration of herdsmen, horses can accurately locate water sources and pastures, ensuring the dynamic balance between nomadic production and grassland recuperation. Horses are not only production tools but also an important part of the no-

madic ecological cycle. At this moment, the wild horse without reins and the text in the subtitle form a complete meaning, representing a complementary relationship at a formal level. Here, they form an expansion relationship within the reinforcing relationship. Expansion refers to one mode providing additional information to supplement and expand the content conveyed by another mode. The subtitles have expanded the meaning of "horse", and the free-roaming horses not only demonstrate the importance that the people of Inner Mongolia attach to animals, but also showcase their nomadic wisdom.


**Figure 5.** The galloping horses.

### 4.3. Venerator of Natural Sanctity

The people of Inner Mongolia hold nature in awe and believe that "true progress is to learn to be humble in the face of nature." Such awe is expressed through specific attitudes conveyed in an interactive meaning. These attitudes can be constructed from the shooting perspective used in images or pictures (Guo & Zhang, 2024). The shooting perspectives can be divided into horizontal, top-down, and bottom-up views. Different angles can highlight the viewer's varying attitudes toward the elements within the image and reflect the power dynamics between them. The documentary shows herdsman Jirgatul splashing the first delicious bite onto the ground when preparing meat for one of Inner Mongolia's classic dishes (**Figure 6**). At this moment, the camera adopts a bottom-up perspective, reversing the conventional power dynamic (humans usually look down at the land), allowing the land to occupy two-thirds of the frame. This bottom-up perspective expresses ultimate reverence for nature and establishes nature as the dominant force. At this time, the linguistic mode "praying" complements the picture, forming two modes that constitute a coordinated relationship without reinforcement in complementary relationship. In the coordinated relationship, different modes work together to completely express the full meaning, and each mode is indispensable. The text supplements the rituals in the above picture and jointly constructs Inner Mongolia's natural ethics of awe and reverence for nature. In addition, at 3 minutes and 47 seconds in the first episode, Jirgatul said, worrying that too many livestock would turn the grassland back into the sandy land again. In systemic functional grammar, the cohesion of textual function involves the logical connec-

tion of information in the text, which ensures the coherence and consistency of the text. This sentence establishes a causal relationship by showing that "too many livestock" is the cause of "pasture degradation." This causal relationship emphasizes the impact of human grazing on the natural environment, highlights the direct effect of human activities on the natural environment, and underscores the land ethics of nomads by strictly adhering to the ecological principle of not exceeding the carrying capacity of pastures.


**Figure 6.** Jirgatul splashed food on the ground.

## 4.4. Interpreter of Ethnic Ecological Aesthetics

In the first episode of the documentary (04:20-04:56), the vast grassland in the background extends to the horizon, forming a vector pointing into the distance, with the distant horizon as its target. This visual structure constitutes the action process of narrative representation. The boundless grassland and the posture of pastoralists sitting quietly on it together form a narrative scene, reflecting the deep connection between the people of Inner Mongolia and the land. It also suggests the relationship between the intergenerational inheritance of Mongolian culture and harmonious coexistence with nature. At this point, the background music is the classic grassland track "Hulunbuir Grassland", played on the morin khuur (horsehead fiddle). The andante tempo at approximately 80 BPM artfully mirrors the natural rhythms of nomadic life. The instrument's distinctive timbre, characterized by resonant overtones produced from horsehair strings, acoustically embodies the Mongolian worldview of ecological harmony. Simultaneously, the melody's ascending fourth intervals sonically amplify the expansive visual vectors of the grassland landscape. Here, an expansion relation within reinforcement in complementary relationship from formal level is formed between the audio and visual modes. In this case, the visual mode has always been at the core of the foreground, assuming the main ideographic function; the auditory mode expands upon the emotion and atmosphere of the visual imagery, rendering the beauty of the grassland's natural scenery—its vast green expanse, the tranquil and crystal-clear lake, the warm-toned sunset afterglow, and the melodious sound of the fiddle—all of which depict the harmony between humanity and nature.

The framing in compositional meaning refers to the actual cropping frame or

spatial dividing lines within an image (Li, 2003). Elements in the image indicate their connection or separation through the most prominent framing lines, color contrasts, etc. In Episode 5 at 01:13 of the documentary, the silhouettes of two Bokh wrestlers overlap against a sunset sky, with deliberate blank space between them. Their intertwined limbs form a tightly framed composition, vividly showcasing the sport of Bokh while projecting the Mongolian cultural ethos of resilience, martial spirit, and integrity. The subsequent subtitle—"Many wrestling moves were developed from the movements of the lion, eagle, and deer"—provides contextual background for Bökh techniques, establishing a non-reinforcing coordinational relation in complementary relationship between modalities. The coordination of textual modality and visual modality constitutes the significance of the Bökh. From a textual metafunction perspective, the clause-initial Theme "Many wrestling moves" immediately anchors the informational focus. The Rheme unfolds through a triad of parallelism ("lion, eagle, and deer"), generating a juxtaposed sequence of steppe creatures. This parallel structure not only creates rhythmic cohesion but also elevates the animals to "masters of technique". Through the Theme-Rheme progression, the ethical principle of "learning from nature" is seamlessly woven into the discourse. Through the imitation of animals, the grassland people transformed their observations of living creatures into respect for them, embodying the unique ethnic aesthetics of Inner Mongolia.

### 4.5. Perserverer of Green Ecology

In practicing the concept of green ecology, the documentary is realized through the attitude in the interactive meaning. In **Figure 7** & **Figure 8**, the lens uses a bottom-up perspective to closely capture Jirgatul's hand, expressing admiration for him. Then the perspective shifts to a top-down view, placing the viewer in a high position and offering an omniscient perspective. This allows the audience to grasp the overarching mission of desert governance at a macro level—an arduous task. The vast desert landscape appears immense but seems to diminish in scale when viewed from above, while the power of an individual's palm appears small from a downward angle yet grows stronger when seen from an upward perspective. **Figure 7** presents Jirgatul's inner monologue, (sand) if not, it likes an evil enemy. From the perspective of the ideational function, participants include "the people of Inner Mongolia" and "the natural environment" (sand). As actors, the people of Inner Mongolia have actively taken measures to control the sand. Sand is used as an object that affects the ecological environment. The "dialogue" here not only refers to linguistic exchanges but also covers the protection and governance actions of the people of Inner Mongolia toward the natural environment. "Enemy"—this sentence personifies the sand in the natural environment, emphasizing the fragility of the ecological environment and the importance of environmental protection. This kind of environmental setting highlights the efforts and achievements of the people of Inner Mongolia in ecological conservation. The textual and visual modalities together create an integrated meaning, establishing an

expansive reinforcement relation in complementarity relation. The text expands the meaning of sand. The contrast in perspectives and the multimodal expansion of text and visuals jointly illustrate the diligent and hardworking people of Inner Mongolia, reflecting Inner Mongolia's determination to practice the concept of green ecology and promote the construction of ecological civilization.


Figure 7. Jirgatul's hand.

Figure 8. The desert land.


Figure 9. The report on the Achievements of Desert Control.

Figure 9 first provides the background of tree planting in the desert. The narration text "Nearly 5 million hectares of farmland and 10 million hectares of grazing land have been preserved by the forest network project. Desertification of around 17 million hectares of land and soil erosion of nearly 13 million hectares of land have been curbed". The transitivity system refers to a semantic framework

through which language represents experiential reality, configuring participants and circumstantial elements via six process types (material, mental, relational, etc.). Within this system, the Material Process denotes actions or events involving energy transfer, where the core participant Actor initiates the action, and the Affected Participant encompasses both the Goal (entity acted upon) and the Range (domain defining the action's scope). Crucially, the Range functions as a non-agentive entity, specifically designating the sphere of impact rather than an object undergoing change. The transitivity system within the ideational function, the passive material processes "have been preserved" and "have been curbed" background human agency by positioning the forest network project as the implicit Actor, thereby highlighting governance outcomes rather than anthropogenic intervention and imbuing ecological conservation with an epistemological necessity. The four sets of spatially quantified data ("5 million hectares of farmland, 10 million hectares of grazing land, 17 million hectares of desertified land, and 13 million hectares of eroded land") function as Range circumstantial elements, concretizing governance scale through geographical metrics. The passive voice strategy syntactically places the land in the position of Affected Participant, grammatically encoding the land as a living entity requiring stewardship—a construction resonating with deep ecological ethics. Collectively, these linguistic mechanisms delineate Inner Mongolia's commitment to the ecological principle that "Green development is the defining feature of high-quality development". The desert in the background gradually turns green, and the visual and linguistic modes form a combinational relation of non-reinforcement in complementarity relation. Combination refers to different modes cooperating with each other to construct a complete meaning. It shows the effectiveness of Inner Mongolia in practicing the concept of green ecology and adhering to the control of desertification. In future development, Inner Mongolia will also adhere to the organic unity of people's wealth and ecological beauty, and solve historical environmental problems.

### 4.6. Practitioner of Sustainable Development

The improvement of the ecological environment has also enhanced the living standards of the people in Inner Mongolia. In the second episode, from 04:26 to 04:49, voice and visual modes form a combinatorial relation within a non-reinforced relation in complementarity relation. In the picture, Jirgatul and his relatives and friends are gathered together. The sequence begins with slow, melancholic piano chords (≈60 BPM) in a minor key, their sustained resonance mirroring the initial challenges of desertification control. As the scene transitions to show improved living conditions, the musical arrangement shifts to brighter major-key arpeggios (≈120 BPM) with higher-register piano tones, acoustically symbolizing renewal. The evolving timbre of the piano audibly traces that through desert management, farmers and herdsmen have developed a harmonious relationship with the ecosystem, transitioning from weighted lower octaves to sparkling upper registers. Higher-grade crops can now be grown, and the production

of beef and mutton has also increased. Together, these two modes convey to the audience both the hardships of sand control and the joy of success, illustrating how hard work has improved the living conditions of the people of Inner Mongolia. By transforming deserts into fertile lands, the region has become a living testament to the idea that protecting nature is investing in the future. The documentary's multimodal storytelling—through its visual models and auditory models, positioning Inner Mongolia as a beacon of sustainable development where lucid waters and lush mountains truly translate into shared prosperity.

This section examines how the documentary constructs Inner Mongolia's ecological identity through multimodal strategies. The film employs high-modality visuals and dynamic vectors to showcase pristine landscapes, while linguistic analysis reveals how material processes and relational clauses encode ecological values. Through bottom-up perspectives and ritual depictions, it establishes nature's sacred status. The documentary effectively combines audiovisual elements-like horsehead fiddle music enhancing grassland scenes-with quantified governance achievements to demonstrate ecological progress. These multimodal techniques work synergistically to present Inner Mongolia as a model of ecological civilization, where cultural traditions, environmental protection and sustainable development coexist harmoniously. It should be noted that the theoretical framework of Visual Grammar relies heavily on structural parallels between visual and linguistic representations, a limitation that may introduce analytical subjectivity during interpretation (Pan & Li, 2017).

## 5. Conclusion

This study demonstrates how *Charming Inner Mongolia* strategically employs multimodal resources to construct a comprehensive ecological image. Through the integrated application of Visual Grammar, Systemic Functional Linguistics, and Zhang Delu's multimodal discourse analysis framework, the analysis reveals six core dimensions of Inner Mongolia's ecological representation: as guardian of pristine ecosystems, transmitter of nomadic wisdom, venerator of natural sanctity, interpreter of ethnic aesthetics, perserverer of green ecology, and practitioner of sustainable development. Ultimately, this study contributes to ecological communication theory by demonstrating how multimodal discourse can bridge the gap between environmental policy, cultural heritage, and public perception-offering a replicable model for sustainable regional development narratives worldwide. The documentary's strategic use of bottom-up camera angles to visually encode reverence for nature (Section 3.3) offers a replicable technique for media producers aiming to highlight indigenous ecological values in other contexts. The effective pairing of high-modality visuals (vibrant landscapes) with absolute linguistic expressions like "great forest" (Section 3.1) demonstrates how reinforced intermodal relations can amplify ecological authenticity in regional branding campaigns. Furthermore, the integration of traditional auditory elements (e.g., horsehead fiddle music) with visual vectors (Section 3.4) provides a model for anchor-

ing cultural identity through multisensory storytelling—a technique particularly valuable for heritage tourism promotions. These replicable models could be applied by other creators of regional promotional media. It is anticipated that subsequent studies will offer more holistic representations of Inner Mongolia's multidimensional identity.

## Funding

This study has been funded by the Basic Research Fund Project for Universities of Inner Mongolia Autonomous Region (No. 20900-54220391).

## Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

## References

Guo, B. Y., & Zhang, D. (2024). Research on the Identity Construction of Luoyang Museum from the Perspective of Visual Grammar. *Journal of Henan University of Science & Technology (Social Science), No. 42,* 103-110.

Halliday, M. A. K. (1973). *Explorations in the Functions of Language.* Edward Arnold.

Halliday, M. A. K. (1994). *An Introduction to Functional Grammar* (2nd ed.). Arnold.

Halliday, M. A. K., & Mattehiessen, C. M. I. M. (2004). *An Introduction to Functional Grammar.* Edward Amold.

Hu, Z. L. (2017). *Introduction to Systemic Functional Linguistics.* Peking University Press.

Jiang, L. X., Zhu, H. H., & Lu, T. H. (2006). Conceptual Analysis of Regional Image and Its Marketing Framework. *Journal of Sun Yatsen University (Social Science Edition), No. 5,* 111-116, 143.

Kress, G., & Van Leeuwen, T. (2021). *Reading Images: The Grammar of Visual Design.* Routledge. https://doi.org/10.4324/9781003099857

Li, H. R. (2019). *The Documentary "Guangxi Stories" Conducts a Study on the Image Construction of Guangxi.* Master's Thesis, Guangxi University.

Li, Z. Z. (2003). Social Semiotic Approach to Multimodal Discourse. *Foreign Language Research, No. 5,* 1-8, 80.

Pan, Y. Y., & Li, Z. Z. (2017). A Comprehensive Review of Multimodal Discourse Analysis in China (2003-2017): Taking the Published Achievements in CSSCI Source Journals as the Research Objects. *Journal of Fujian Normal University (Philosophy and Social Sciences Edition), No. 5,* 49-59, 168-169.

Tian, H. L., & Zhang, X. J. (2013). Meaning in Images and the Ideology of Media: A Multimodal Discourse Analysis Perspective. *Foreign Language Research, No. 2,* 1-6.

Van Ieeuwen, T. (2005). *Introducing Social Semiotics.* Routledge.

Wei, Q. H. (2008). On the Overall Meaning Construction of Multimodal Discourse: Discourse Analysis Based on a Multimodal Media Discourse. *Journal of Tianjin Foreign Studies University, No. 6,* 16-21.

Yuan, X. L., & Nai, R. H. (2022). Research on the Construction of Multimodal Discourse Meaning in the International Communication of Cultural China. *Foreign Language Education, No. 5,* 23-29.

Zhang, D. L. (2009). Exploration of the Comprehensive Theoretical Framework for Multi-

modal Discourse Analysis. *Foreign Languages in China, No. 1,* 24-30.

Zhang, D. L., & Wang, Q. (2011). The Graphic Relationship and Interpretation Process of Traffic Signs. *Foreign Language Education, No. 4,* 27-30, 35.