A Multimodal Discourse Analysis of the Documentary *Discover China: Behind the Rapid Growth*

Yuhan Zhao, MA
Ocean University of China, China

Doi: 10.19044/llc.v12no1a7 https://dx.doi.org/10.19044/llc.v12no1a7

Submitted: 23.07.2025
Accepted: 21.11.2025
Published: 05.12.2025

Cite As:
Zhao, Y. (2025). *A Multimodal Discourse Analysis of the Documentary Discover China: Behind the Rapid Growth.* International Journal of Linguistics, Literature and Culture, LLC, 12 (i-1), 130. https://dx.doi.org/10.19044/llc.v12no1a7

Abstract

With the growing significance of visual and auditory elements in the digital age, multimodal discourse has emerged as a critical area of study. Drawing on Kress and van Leeuwen's Visual Grammar and Zhang Delu's multimodal discourse analysis framework, this study investigates the documentary titled Discover China: Behind the Rapid Growth, focusing on how representational, interactive, and compositional meanings are realized through multimodal resources. The analysis is based on 28 representative shots from the episode centered on Yiwu, utilizing ELAN 5.9 software to annotate and examine various semiotic modes including prosody, image, animation, and verbal expression. Quantitative and qualitative analyses reveal that diverse modes interact primarily in complementary ways to convey national narratives and promote China's image. The findings highlight the dynamic interplay among modalities, with linguistic narration supported by visual salience and auditory emphasis. This study contributes to the expanding field of multimodal discourse by offering empirical insights into dynamic audiovisual texts and provides a model for analyzing promotional documentaries from a linguistic and cultural perspective.

Keywords: Multimodal discourse analysis; visual grammar; Zhang Delu; documentary; China image

Introduction

With the technological development and the popularity of the internet, communication has increasingly relied on multiple semiotic resources rather than on language alone. While language continues to serve as the primary

medium of meaning-making, it is complemented by other modes such as images, sounds, colors and gestures. The coexistence and interaction of these different modes in communication constitute what is generally termed multimodality. When these semiotic resources are combined in specific communicative practices---such as movie posters that contain both text and images, conversations that contain variations of intonation and gestures, and TV advertisements or documentaries---they give rise to multimodal discourse. Multi-modal Discourse Analysis (MDA), therefore, according to Wei (2009), refers to is the analysis of several or all of the different semiotic modes in a text or communicative event.

This study chooses a documentary released by *China Daily* named *Discover China: Behind the Rapid Growth* as the object of analysis because it represents an example of a country-image documentary in Chinese media and provides rich material for multimodal analysis. The analysis is conducted using ELAN 5.9, a professional software for analyzing multimodal discourse, based on the Visual Grammar proposed by Kress & van Leeuven (1996) and MDA comprehensive theoretical framework by Zhang Delu.

The significance of this study mainly covers the following two aspects: Firstly, the study expands the scope of research material for MDA. Li (2023: 2) notes that most existing studies focus on films, advertisements, city-image promotional films, teaching cases, or posters, while country-image documentaries are rarely examined. Second, by providing a detailed multimodal analysis of this documentary, the study may offer useful insights for documentary creators who aim to construct and communicate a comprehensive and positive image of China internationally.

Over the past decades, scholars have come to recognize that interpreting discourse solely through language is inadequate: other semiotic resources-such as images, diagrams, music, gesture, and spatial layout-play active roles in the construction of meaning. Consequently, the study of multimodality has developed into a distinct field of inquiry. The term multimodality denotes the coexistence and interaction of multiple semiotic systems (Gu, 2015), and multimodal discourse refers to communicative events in which meaning arises from the coordinated deployment of linguistic, visual, auditory and other modalities (Li, 2003; Baldry & Thibault, 2006).

Foundational contributions from semiotics and systemic functional linguistics have strongly shaped multimodal discourse analysis (MDA). Roland Barthes' early semiological work on the relations between image and text (Image–Music–Text, 1977) provided conceptual tools for thinking about interplay among modes, even though it emerged from the semiology tradition rather than from MDA proper. Building on Halliday's

metafunctional perspective, scholars such as O'Toole and Kress & van Leeuwen extended systemic functional ideas to visual meaning: Kress and van Leeuwen's Reading Images (1996) proposed a visual grammar that parallels linguistic metafunctions and has become central to visual analysis; their subsequent work on colour (2006) further explored how non-linguistic resources can be treated as semiotic systems. Bateman (2008) and O'Halloran (2004, 2011) contributed both theoretically and methodologically, with O'Halloran notably articulating frameworks for film discourse and developing platform-oriented tools to support model-based multimodal analysis.

In recent years the international literature has broadened in two notable ways. First, methodologically, there has been increasing attention to dynamic and digital genres-video, social media, and interactive platforms-and to analytic techniques that combine qualitative description with computational or visualization tools (see Jewitt, 2016; Chen, 2019). Second, empirical work has diversified across domains (education, climate communication, news media), showing how modal interactions vary according to platform affordances and temporal patterns (e.g., studies of audio description and multimodal literacy; and recent studies that examine dynamic social-media multimodality and large multimodal corpora). These developments underscore the field's shift from static image/text analysis toward the study of temporally unfolding, digitally mediated multimodal practices (Wang & Taabaldiev, 2025; CliME project, 2025).

China has an active tradition of MDA scholarship that both imports and indigenizes international theories. Early introductions of visual grammar and SFG-informed analysis (e.g., Li Zhanzi, 2003; Hu, 2007; Zhu, 2007) laid the groundwork for subsequent methodological applications in pedagogy and media studies. Zhang Delu's four-level framework (2009)-comprising cultural, contextual, content and expression levels-offers a systematic model that connects macro-sociocultural conditions with micro-level modal resources and their formal realization; Zhang further distinguishes modal relations as complementary or non-complementary, a distinction adopted as a key analytical lens in the present study. Subsequent Chinese research (e.g., Zhang & Yuan, 2011; Zhao & Feng, 2017; Li & Feng, 2017) has applied and extended these frameworks to dynamic broadcast genres and to ideological analysis.

Taken together, the international and Chinese literatures provide robust theoretical and methodological resources for multimodal analysis. Nevertheless, two gaps remain. First, many domestic studies-while theoretically sophisticated-have focused on relatively static multimodal texts (e.g., posters, print adverts, still images) or on pedagogical applications; fewer have systematically examined dynamic documentary genres in which

modalities interact over time. Second, although international work increasingly addresses digital, temporal and computational approaches, there remains a need to synthesize these methodological advances with the culturally grounded, SFG-informed frameworks developed within Chinese scholarship. This study therefore positions itself at the intersection of these strands: it adopts Zhang's four-level framework as the theoretical foundation while drawing on recent international methodological advances to analyze the temporal coordination of language, sound and image in documentary discourse. (Methodological choices-e.g., the use of ELAN for time-aligned annotation-are explained in detail in the Methodology chapter.)

Documentary Analysis from the Perspective of Multi-modal Discourse

Documentaries, as a good research material that contain many sorts of semiotic modes, such as actions, colors, images, and languages, have attracted the interest of linguistics. Zhang and Yang (2021) aim to delve into the narrative strategy and meaning construction in the television documentary titled *China Women's Volleyball Team*. Their analysis focuses on both the presentation techniques and discourse expressions employed in the documentary. Furthermore, utilizing a multimodal discourse analysis framework, they explore the cultural inter-construction and interpretation generated within the documentary from an "audience-text" perspective.

Duan (2019) tries to make an in-depth multimodal analysis of *Amazing China*. By the Visual Grammar of G. Kress and van Leeuwen, she analyzes how the verbal modes achieve the construction of discourse meaning and how these modes cooperate in the construction of meaning. Through Zhang Delu's complementary relationship theory, she tries to analyze the relationship between the modes.

Liu and Zhang (2018) analyze multimodality to provide some useful insights for the optimization and reshaping of China's national image. In addition, Tang (2016) makes multimodal discourse analysis of the BBC documentary *Beautiful China*, and Liang (2018) makes multimodal discourse analysis of the BBC documentary *Chinese New Year: The World's Biggest Feast*.

To summarize, Multimodal Discourse Analysis has obviously become a hotspot. Compared with the earlier static discourse analysis, current research is increasingly turning to the dynamic ones. However, as a vibrant discourse, though there is a little research about documentary, it has not received enough attention. Therefore, this paper chooses the documentary on publicity abroad made by the mainstream media *China Daily* as the subject of study, *Discover China: Behind the Rapid Growth*, and analyzes its content from the perspective of multimodal discourse.

Theoretical Framework

Two main theories will be introduced in detail, which are Visual Grammar Theory of Kress & van Leeuven and Zhang Delu's comprehensive theoretical framework of multi-modal discourse analysis.

Visual Grammar Theory

Halliday's Systemic Functional Grammar views language as a social symbol. In 1985, he proposed three meta-functions of language: ideational function, interpersonal function and textual function. Ideational function refers to the role of language in expressing people's experience in the real world and inner world. Interpersonal function expresses the speaker's identity, status, attitude and inference of things, and people use language to attend to social activities and establish social relations. Textual function refers to the function of language itself which is coherent and related to register.

According to Kress & Leeuwen: "Just as the grammar of language determines how words form clauses, sentences and discourses, visual grammar explains how characters, places and things depicted form visual representations of different levels of complexity" (1996). They proposed Visual Grammar Theory in 1996 on the basis of Halliday's three meta-functions of SFG. They defined that the three meta-functions in SFG correspond respectively to the three meanings in VG: Representational meaning, Interactive meaning and Compositional meaning. And the realization of the three meanings is through a set of dimensions, which can be seen from the following graphic.

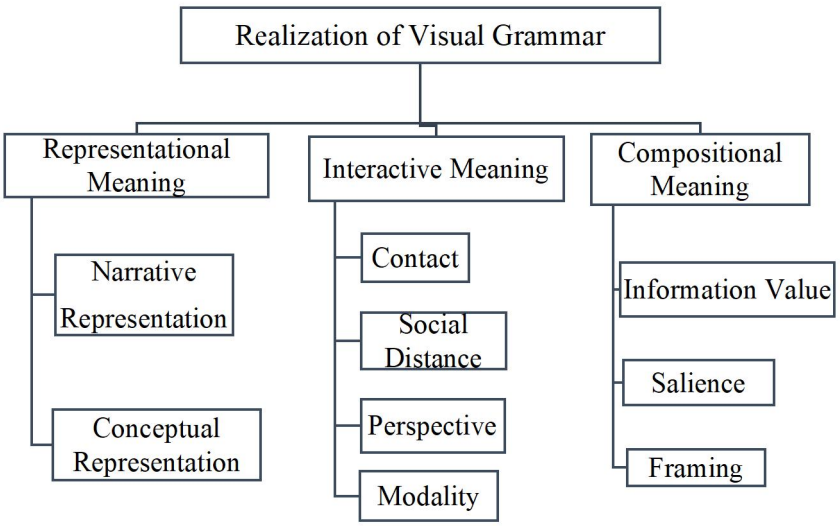

Figure 1. The Realization of Visual Grammar

The representational meanings of image correspond to the ideational function of language in SFG. It refers to people's intuitive feeling about representation of things depicted in visual image, and can be divided into two types, namely "narrative representation" and "conceptual representation". The distinction between the two types is that there is a "vector" in narrative representation but not in conceptual representation. Vector is the direction of action, behavior, usually a slash. When participants are connected by "vectors", they are reproduced as doing something to each other. Narrative images show actions and events in development, process of change, and transitory spatial arrangement. Representation of visual structure can be conceptual that represent "participants in terms of their more generalized and more or less stable and timeless essence, in terms of class, or structure or meaning" (Kress & van Leeuwen, 1996).

Interactive meanings of image in visual grammar corresponds to the interpersonal function of language in Systemic Functional Grammar. It explains how images interact with the viewer. Images can be analyzed on four levels, namely contact, social distance, perspective and modality (Kress & van Leeuwen, 1996).

The compositional meaning of visual grammar corresponds to the textual function of SFG. The compositional meaning refers to how the image integrates the representational meaning and interactive meaning to form a meaningful whole.

Zhang Delu' Analysis Framework of Multi-modal Discourse

Zhang Delu (2009) proposes a systematic framework for multimodal discourse analysis that has become an influential theoretical foundation in this field. The framework consists of four interrelated levels-cultural, contextual, content, and expression-which together explain how meaning is constructed across different semiotic resources. At the cultural level, multimodal discourse is shaped by ideology, social consciousness, value systems, and genre conventions, which provide the basis for communicative practices and modality choices. The contextual level, drawing on Halliday's concepts of field, tenor, and mode, emphasizes the influence of situational factors on how semiotic resources are combined. The content level integrates both meaning and formal dimensions: it encompasses ideational, interpersonal, and textual meanings as well as the formal systems specific to each modality, such as linguistic grammar, visual grammar, and auditory organization. The expression level refers to the material realization of discourse, including both linguistic and non-linguistic media such as speech,

writing, font, gesture, image, sound, and digital platforms, which transform abstract meanings into perceivable forms.

Within this four-level framework, Zhang also highlights the relations between modalities, which can be either complementary or non-complementary. Complementary relations are more common in multimodal communication, occurring when one modality cannot fully convey meaning on its own and requires the support of another. These relations can take the form of reinforcement, in which one mode dominates while another enhances it, or non-reinforcement, in which multiple modalities jointly construct meaning and none can be omitted without loss. Non-complementary relations, by contrast, occur when the additional modality contributes little new information but still makes the discourse more vivid, concrete, or accessible. For example, statistical data expressed through speech may be accompanied by charts that, while not adding fundamentally new content, enhance clarity and precision.

By combining the four-level framework with the distinction between complementary and non-complementary relations, Zhang provides a comprehensive model that accounts for both the structural organization of multimodal discourse and the dynamic interplay among its constituent modalities. This study adopts Zhang's framework as its theoretical foundation, as it enables both macro-level analysis of socio-cultural and contextual influences and micro-level investigation of modality interactions in documentary discourse.

To sum up, this study adopts visual grammar as the theoretical basis to analyze how the documentary titled *Discover China: Behind the Rapid Growth* realizes representational, interactive, and compositional meanings. In addition, Zhang Delu's framework of multimodal discourse relations is applied to examine how different modes interact within the documentary.

Methods

This section focuses on research methodology, which probes into the way the study is designed and conducted. Research questions are given first, and it introduces the research subject, that is the chosen discourse documentary *Discover China: Behind the Rapid Growth.* Then the research instrument and procedures are presented.

Research Questions

This study attempts to find answers to the following three questions:

1) What are the distribution features of visual and auditory modes in *Discover China: Behind the Rapid Growth*? Is the frequency, percentage and length of each mode influential to the presentation of the documentary theme?

2) How are the representational, interactive and compositional meanings realized in *Discover China: Behind the Rapid Growth*?
3) What is the relationship between Multi-modal modes in *Discover China: Behind the Rapid Growth*?

Research Subject

Discover China: Behind the Rapid Growth is an 18-episode documentary series, each episode lasting about 6-10 minutes, released by *China Daily* to celebrate the 40th anniversary of reform and opening up. Three foreign journalists explore China's 40 years of reform and opening up as listeners. With the question of "what changes have taken place in China in the past 40 years" in mind, they set off from Shenzhen, which used to be a small fishing village and traveled to the Guangdong-Hong Kong-Macao Bay Area, the Yangtze River Economic Belt, and Beijing-Tianjin-Hebei region in three different directions, as well as the Western Development Area and the rising city clusters in central China, and finally met in Xiong'an. Through the perspectives of foreign journalists, the documentary shows the international community the brilliant achievements in various economic and social fields over the past 40 years of China's reform and opening-up. It embodies and highlights the culture and image of modern China.

Due to limitations of time and resources, this paper selects the ninth episode as the research corpus, which narrates about Yiwu, a city famous worldwide for its wholesale markets for small commodities in Zhejiang province after the reform and opening up. The reason why the author chooses this episode is that Yiwu City epitomizes the development of China after 1978. And 28 shots are chosen to be analyzed and discussed in the study.

Research Instrument

ELAN (EUDICO linguistics Annotator) software is a convenient and practical multimodal corpus analysis tool developed by Max Planck Psycholinguistics Institute in the Netherlands, which provides free technical support for language analysts. Lan (2020) shows that ELAN can achieve multi-level synchronous annotation of video files, including discourse content, voice tone, facial expressions, hand movements, etc. Wang and Wen (2008) summarize the following advantages of ELAN software: video playback is accurate to 0.01 seconds; annotation is synchronized with text, sound and image; the annotation layer is infinite, and the annotation coding table is self-defined; the annotations are linked to each other, and the annotation results can be output in multiple formats according to the research requirements.

With the help of multimodal discourse analysis software ELAN 5.9, the documentary is transcribed, and the frequency of various modalities is counted so as to analyze the different elements in the representational meaning, interactive meaning and compositional meaning.

Research Procedures

Quantitative analysis and qualitative analysis are both used in this study. Firstly, with the help of a popular video software, ELAN 5.9, a small corpus with the 28 shots was created. Secondly, this study chooses some specific shots to analyze the realization of representational meaning, interactive meaning and compositional meaning. Thirdly, an exploration is made about the relationship between the main forms of modalities.

Results and Discussion

This part firstly uses the method of quantitative analysis, from the expression aspects, with the help of ELAN 5.9, to transcribe and annotate an episode of the documentary Discover China: Behind the Rapid Growth. Then, on the basis of Visual Grammar Theory of Kress & van Leeuven, the author uses qualitative analysis to interpret the chosen shots so as to illustrate how the documentary finishes the realization of the three meanings in VG. Last, based on the multi-modal discourse analysis framework proposed by Zhang Delu, the relationships between some main modalities are analyzed to explain how they cooperate to achieve the construction of the whole meaning and the theme of the documentary.

The Expression Aspects

With the help of ELAN 5.9, the author analyses the selected documentary across four dimensions: prosody, image, animation, and expression. Table 1 shows the details of the four aspects respectively.

Table 1. Annotation Patterns

Expression Aspects	Annotation	Meaning
Prosody (P-)	PIF	Prosody-Intonation Fall
	PIR	Prosody-Intonation Raise
	PIRF	Prosody-Intonation Raise/Fall
	PFT	Prosody-Fast Tempo
	PST	Prosody-Slow Tempo
	PSt	Prosody-Stress
	PPS	Prosody-Pause Short
Image (I-)	IP	Image-Picture
	IT	Image-Text
	ITP	Image-Text/Picture
Animation (A-)	AP	Animation-Person
	AT	Animation-Thing
	ATP	Animation-Thing/Person

Expression (ES-)	ESD	Expression-Style Dialogue
	ESM	Expression-Style Monologue
	ESN	Expression-Style Narration

The documentary is divided and labeled according to the above annotation patterns. It can be seen in the labeled file (Figure 2) that it is divided into 7 levels, namely: Intonation, Tempo, Stress, Pause, Image, Animation, Expression. The author repeatedly views and compares the segmented and hierarchically labeled videos, and exported the labeled data for further quantitative analysis. See the result in Table 2.

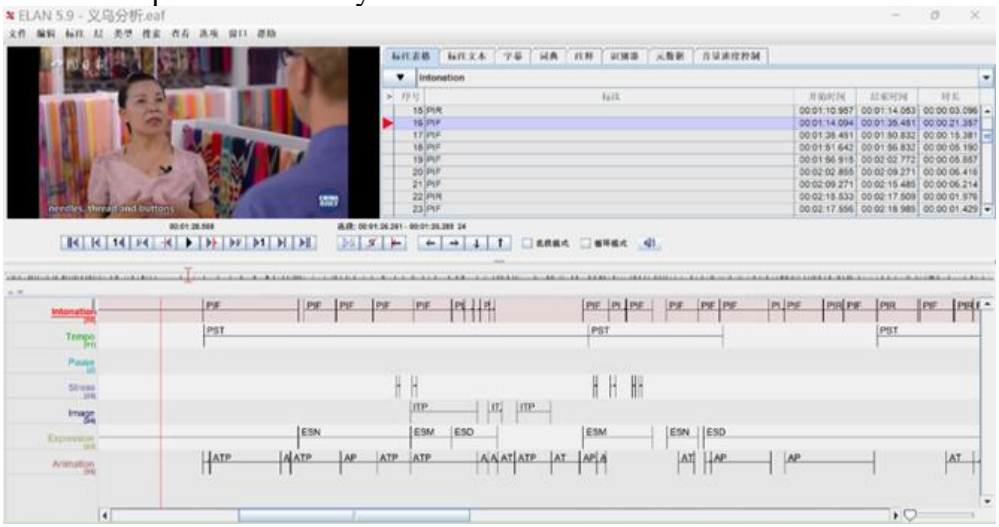

Figure 2. The Labeled File

Table 2. The Distribution of Prosody Modality

Annotation	PIF	PIR	PIRF	PFT	PST	PSt	PPS
Occurrences	71	17	2	6	5	20	2
Minimal Duration (seconds)	0.381	0.560	1.857	5.810	2.048	0.357	1.905
Maximal Duration (seconds)	21.357	6.667	4.285	58.996	62.000	1.178	2.012
Average Duration (seconds)	5.461	2.600	3.071	27.747	22.601	0.674	1.959
Median Duration (seconds)	4.881	2.571	3.071	24.563	19.848	0.608	1.958
Total Annotation Duration (seconds)	387.723	44.206	6.142	166.485	113.005	13.470	3.917
Annotation Duration Percentage(%)	81.289	9.268	1.288	34.905	23.692	2.824	0.821
Latency	0.000	14.255	378.612	39.987	60.326	2.460	38.053

Table 2 indicates that falling tones (PIF) dominate both frequency (71 instances) and cumulative duration (387.7s, 81.3%), whereas rising and level tones are marginal. This distribution highlights a prosodic preference for closure and finality. Within Halliday's interpersonal metafunction, such tones function as rhetorical resources for projecting authority and certainty, reinforcing the documentary's didactic style. Latency, defined as the interval between successive prosodic units, reflects the rhythm of prosodic transitions; longer latencies mark topic shifts. These findings address RQ1, showing how prosodic cues contribute to an authoritative auditory frame, which complements the visual narrative of national development.

Table 3. The Distribution of Image Modality

Annotation	IP	IT	ITP
Occurrences	2	1	17
Minimal Duration (seconds)	2.112	0.778	1.361
Maximal Duration (seconds)	4.500	0.778	10.820
Average Duration (seconds)	3.306	0.778	3.813
Median Duration (seconds)	3.306	0.778	2.739
Total Annotation Duration (seconds)	6.612	0.778	64.813
Annotation Duration Percentage(%)	1.386	0.613	13.589
Latency	284.540	388.640	8.460

As shown in Table 3, integrated text–picture sequences (ITP) are predominant (17 instances; 64.8s, 13.6%), compared to isolated text or images. Quantitatively, this indicates reliance on multimodal composites rather than stand-alone visuals. In Kress and van Leeuwen's visual grammar, such integration exemplifies representational complementarity: verbal anchorage provides explicit propositions, while visuals contextualize them. The strategy enhances coherence and accessibility, contributing to RQ2 by demonstrating how multimodal resources jointly reinforce meaning-making.

Table 4. The Distribution of Animation Modality

Annotation	AP	AT	ATP
Occurrences	20	36	40
Minimal Duration (seconds)	1.120	0.429	1.071
Maximal Duration (seconds)	14.405	8.796	20.560
Average Duration (seconds)	4.005	3.026	5.053
Median Duration (seconds)	2.679	1.797	3.229
Total Annotation Duration (seconds)	80.105	108.947	202.130
Annotation Duration Percentage(%)	16.795	22.842	42.378
Latency	0.850	3.250	2.000

Table 4 shows that combined animations of person and thing (ATP) account for the largest share (40 occurrences; 202s, 42.4%), followed by object- and person-only animations. This quantitative pattern reflects a

preference for visualizing human–object interaction over abstract or isolated representations.

Table 5. The Distribution of Expression Modality

Annotation	ESD	ESM	ESN
Occurrences	8	9	8
Minimal Duration (seconds)	7.379	6.440	5.077
Maximal Duration (seconds)	56.000	23.195	27.110
Average Duration (seconds)	26.625	11.641	14.790
Median Duration (seconds)	19.943	11.261	15.416
Total Annotation Duration (seconds)	213.004	104.771	118.317
Annotation Duration Percentage(%)	44.658	21.966	24.806
Latency	60.680	8.660	0.000

Dialogues dominate the expressive modality (213s, 44.7%), with monologues and narration playing secondary roles. This polyphonic distribution highlights the documentary's emphasis on diverse social voices. Within Zhang's framework, dialogue and narration form a complementary relation: narration ensures continuity, while dialogue injects immediacy and authenticity. In discourse terms, this balance avoids monotony, sustains audience engagement, and reinforces credibility. These findings respond to RQ3, showing how expressive choices serve the persuasive construction of China's development narrative.

The Contents Aspects

After the quantitative analysis, the author uses qualitative analysis to interpret how the documentary realizes the three meanings of Visual Grammar.

The Realization of Representational Meaning

The representational meaning can be divided into narrative representation and conceptual representation, the difference is that the former contains a "vector", whereas the latter does not To be more specific, the narrative representation can be further divided into three types, which are action process, reaction process, speech and mental process. Usually there are two participants in the image involved in action process, namely "Actor" and "Goal". The "vector" is embodied by objects in action taken by "Actor" and directed at "Goal".

Fig. 3: An interview between the host and a Senegalese businessman

The use of this image adheres to the Fair Use Agreements in Copyright Law.

Figure 3 is taken from an interview between the host and a Senegalese businessman who has lived in China for over a decade. From a representational perspective, the "vector" is formed by the businessman's hand movement as he lifts and pours from a Chinese teapot toward the host. The action establishes a transactional process in which the businessman acts as the Actor and the host as the Goal. Framing and medium shot composition allow both participants to be visible within the same spatial field, emphasizing their mutual engagement. The teapot, positioned at the center of the frame, functions as a cultural artifact mediating this interpersonal exchange.

When a "vector" consists merely of the direction of the eyes of the participants in one or more diagrams, it is called a reaction process. The following two shots are examples of this process.

Fig. 4: An interview between the host and the first-generation Yiwu vendor

The use of this image adheres to the Fair Use Agreements in Copyright Law.

Figure 4 presents the interview between the host and the first-generation Yiwu vendor. From this shot, the vendor's gaze is clearly visible and directed toward the host, whereas the host's gaze direction remains off-screen. This kind of layout mainly aims to make this vendor a chief narrator of the development of Yiwu City to audience. Although this conversation happens between the host and the vendor, this scene makes the audience feel that they are also in the conversation personally.

Fig. 5: An whole perspective of one interview between the host and entrepreneur.

The use of this image adheres to the Fair Use Agreements in Copyright Law.

Unlike Figure 4, Figure 5 depicts the frame of interview from the whole perspective. The audience can see the eye direction of both the interviewer and the interviewee obviously. From the perspective of Visual Grammar, it is a reactional process. By looking at each other, the two in the shot achieve equilibrium and interaction. The audience can obtain a feeling of equal communication.

According to Kress & van Leeuwen (1996), conceptual process represents relatively stable and timeless participants in general in terms of its classification, structure and meaning.

Fig. 6-8: All sorts of commodities sold in Yiwu City.

Fig. 6

Fig. 7

Fig. 8

The use of these images adheres to the Fair Use Agreements in Copyright Law.

From the above shots, it is not difficult to find that these images are the presentation of classification process. They are all commodities sold in Yiwu City. Through the static display of these merchandises, the meaning that Yiwu City has the reputation of artifact is constructed.

Symbolic process is about what the participants are or mean. It is concerned with the meanings formed by participants of the images.

Fig. 9: A teacup

The use of this image adheres to the Fair Use Agreements in Copyright Law.

In the above shot, there is a teacup. It is prominent in the image and carries the symbolic meaning of Chinese culture. From the video, it is used proficiently by the foreign businessman. This shot illustrates that behind the rapid growth of the Chinese economy, there is also the contribution of Chinese culture.

The Realization of Interactive Meaning

Interactive meaning is about the relationship between the maker of the image, the object represented by the image and the viewer of the image. Because one of the main purposes of documentary makers is to convey to the audience the rapid development of China, it is particularly important to construct interactive meaning. It is mainly realized through four elements: contact, social distance, perspective and modality.

First, contact refers to an imagined interpersonal relationship established by the participants and viewers through eye contact. There are two orientations for contact, “demand” and “offer”. When the participants are living things and have direct eye contact with the viewers, seeming to be asking the viewer for something, it is called “demand”. While “offer” means

that the image where participants do not have eye contact with the viewer, or they are not living things, but convey relevant information to the viewers.

Fig. 10-12: The host is introducing the Belt and Road Railway.

Fig. 10 **Fig. 11** **Fig.12**

The use of these images adheres to the Fair Use Agreements in Copyright Law.

Figures 10-12 are of the host introducing the Belt and Road Railway. From the perspective of Visual Grammar, it is a contact of interactive meaning. To be more specific, it belongs to "demand" because the host makes eye contact with the audience. This kind of positive eye communication is like asking for the audience's understanding and attracting their attention. The three different scenes provide the audience with a relatively panoramic outlook of the Belt and Road Railway so that the audience have the desire for more information and knowledge about it.

Fig. 13-15: The eye contact and facial expression in the interview between host and Yiwu vendor.

Fig.13 **Fig.14** **Fig.15**

The use of these images adheres to the Fair Use Agreements in Copyright Law.

As a documentary introducing China's rapid development, the main purpose of it is to provide audience with information on China's achievements. Therefore, most of the shots belong to the "offer", and only a few of the host's introduction belong to "demand". In figures 13-15, the "offering act" can be inferred from participant's eye contact and facial expression. In figures 13–14, He Haimei-the first-generation Yiwu vendor-is portrayed speaking with the host in a textile shop. Her gaze is directed toward the host rather than the camera, forming an offer structure that presents information rather than seeking direct viewer engagement. The camera adopts a medium shot at eye level, creating a social distance that positions viewers as witnesses to an authentic exchange rather than as interlocutors. The frontal horizontal angle fosters a sense of inclusion, while

the balanced lighting and colorful fabric background visually reinforce her credibility and industriousness. Collectively, these semiotic cues construct He Haimei as an emblem of Yiwu's grassroots entrepreneurship-accessible yet dignified-thereby contributing to the documentary's overarching narrative of ordinary individuals driving national progress. In figure 15, the close-up and slight upward angle enhance intimacy and align the viewer with the host's receptive gaze. According to his facial expression, it seems that the host is listening to He Haimei attentively and carefully. This also implicitly shows how interesting her story is.

The second element related to interactional meaning is "distance", which is associated with the frame size of the camera view. It can allow the audience to be close to the participants, place and events of the image. It usually depends on three elements: close shot, medium shot, and long shot.

Fig. 16-17: The panorama of Hangzhou Bay Bridge and Yiwu part of the Belt and Road Railway.

Fig.16

Fig.17

The use of these images adheres to the Fair Use Agreements in Copyright Law.

Figure 16 shows the Hangzhou Bay Bridge from the perspective of the "long distance", which includes almost the whole appearance. Figure 17 displays the panorama of Yiwu part of the Belt and Road Railway. While presenting the comprehensive and authentic pictures, these shots indicate the relationship between the participant and audience is an objective social distance.

There is another element used frequently in the documentary, which is modality. It includes three types: high modality, medium modality and low modality. Modality in visual grammar indexes the degree of pictorial "truth-value" and is realised through features such as colour saturation, lighting (natural vs. staged), surface detail (texture), depth cues (foreground–background separation), focus/sharpness, and perspectival realism.

Fig. 18-20: The presentation of textiles and goods.

Fig.18 **Fig.19** **Fig.20**

The use of these images adheres to the Fair Use Agreements in Copyright Law.

In **figures 18-20** the film employs high modality: textiles and goods are captured with vivid, saturated colours, strong local contrast and crisp focus (fig. 18), overhead patterned displays create layered depth and dynamism (fig. 19), and warm, specular highlights on decorative lamps produce tactile surface detail (fig. 20). Compared with earlier interview shots (medium modality: neutral lighting, softer contrast, eye-level framing), these market sequences foreground spectacle and materiality rather than interpersonal testimony. The higher visual verisimilitude here functions representationally-making commodities appear tangible and abundant-and compositionally by increasing salience (colour and light draw the eye) and information value (foregrounding goods as the centre of attention). Theoretically, this high-modality treatment complements the narration (a complementary relation in Zhang's terms): while voiceover supplies propositional claims about Yiwu's commerce, saturated, high-detail visuals supply perceptual evidence that bolsters plausibility and elicits desire. Thus, modality is mobilised not merely for aesthetic effect but as a multimodal resource that enhances the documentary's persuasive construction of Yiwu as a vibrant commercial hub.

The Realization of Compositional Meaning

Simply put, compositional meaning refers to the overall layout of the image. According to Kress & van Leeuwen, it involves three elements in the image: information value, salience and framing.

The information value is realized by the placement of elements in the image. The position of the elements in the image, such as left or right, center or margin, top or bottom, gives them different information values.

Fig.21: The host is introducing the Belt and Road Initiative in front of CHINA RAILWAY Express.

The use of these images adheres to the Fair Use Agreements in Copyright Law.

The above shot is of introducing the Belt and Road Initiative, displaying a piece of clear information "中欧班列 CHINA RAILWAY Express". These characters are in the center position of the image, while other elements are marginalized as background. This information layout definitely underlines China Railway Express.

Salience refers to the different degrees of attraction of the elements in the image to the viewers. It can be achieved by perspective, size, color, focus and so on.

Fig.22: The sunrise.

The use of these images adheres to the Fair Use Agreements in Copyright Law.

In figure 22, through the hole of the building, a sunrise can be seen without any obstruction. The dazzling light brought by the sunrise contrasts with the surrounding buildings. And the contrast in color makes the sunrise more eye-catching. This shot realizes meaning with the representation of "the rising sun". It also implies that the development of China is like a "rising sun" and it will have a better prospect in all aspects of economy, science and technology, culture and so on.

Framing refers to the "connecting" and "disconnecting" of the represented participants in the visual images. In this video, there are lots of consecutive "different" shots in the video, but they express continuous and complete meanings.

Fig. 23-26: The host and He Haimei are talking about the old album.

Fig.23

Fig.24

Fig.25

Fig.26

The use of these images adheres to the Fair Use Agreements in Copyright Law.

These shots are taken from the documentary from 00:01:02--00:01:07. They are connected through different shots but form the same events. Figure 23 shows He Haimei flipping through old album with her hands. Figure 24 switches to a closer lens, giving this album a close-up shot and showcasing its vintage feel. Figure 25 gives this album a frontal shot, allowing the audience to see the contents clearly. And figure 26 shows He Haimei talking about this album and the host listening to her carefully. These shots may vary in their content, yet they are unified within the album, conveying a consistent and cohesive narrative. By linking these diverse shots, audience gains a holistic understanding of the entire sequence. Therefore, it is imperative to notice that the connection and disconnection of the represented participants in the visual images work collaboratively to express the whole meaning of the video.

The Relationship Between Different Modes

As the documentary falls within the realm of dynamic multi-modal discourse, it is necessary to emphasize the interplay among various modal forms, including visual, verbal, and auditory modes. Following Zhang Delu's classification (2009), these multi-modal modes and their interactions can be

categorized into two distinct types: complementary and non-complementary relationships.

Complementary Relationship

The concept of a complementary relationship can be further delineated into two categories: reinforced and non-reinforced relationships. In a reinforced relationship, one mode serves as the primary means of communication, while other modes serve to bolster or support it. The documentary demonstrates a significant complementary relationship between auditory and visual modes.

Fig.27: The host is visiting the street food of Yiwu City.

The use of this image adheres to the Fair Use Agreements in Copyright Law.

In the shot above, the audible mode is the host's narrative "Fortunately, Yiwu is known for its street food". This commentary constitutes a comprehensive assessment of the event, providing audience with an overall understanding of the street food in Yiwu. From the visual perspective, the scene is filled with a variety of food stalls bustling with numerous visitors. These elements provide background information for the auditory mode's narrative, reinforcing the significance of the auditory mode's discourse.

In the non-reinforced relationship, various modes collaboratively convey the overarching meanings of the communication, and the absence of any one mode results in incompleteness.

Fig.28: The commodities about "Santa" in Yiwu City.

The use of this image adheres to the Fair Use Agreements in Copyright Law.

In figure 28, the auditory mode is the narrative of the host "Apparently, Santa lives in Yiwu". This narrative uses metaphor. If the audience only has this mode, there may be some confusion. However, with the visual mode which shelves full of Santa toys, the audience can certainly get to know the meaning by the host.

Non-complementary Relationship

The non-complementary relation implies that the second mode adds minimal value to the first mode, yet it remains present as a mode. Even though the second mode does not offer distinct information, it can potentially enhance the clarity and visual representation of the information conveyed by the first mode.

Fig.29: High buildings in Yiwu City.

The use of this image adheres to the Fair Use Agreements in Copyright Law.

In figure 29, the audible mode is the monologue of the host "The sector is booming and is propelled by such companies." Only by this sentence, the audience actually can understand the meaning. But the

producer of the documentary adds a visual mode, which is about the high building of "such companies". By doing this, the meaning conveyed by the audible mode is strengthened though it does not offer any new information.

Conclusions

This chapter consists of three sections: research findings and limitations

Research Findings

This study investigated how multiple semiotic modes interact to construct meaning in the documentary Discover China: Behind the Rapid Growth, combining quantitative annotation data with qualitative visual-grammar analysis. The findings reveal that visual and auditory modes dominate the multimodal configuration, while expressive and linguistic resources provide crucial cohesion. Quantitatively, mixed person–thing animation occupies the largest visual share (42%), dialogue the most frequent expressive form (45%), and falling tones prevail in prosody. Qualitatively, these patterns align with the documentary's reliance on narrative representation, where embodied actions and speech vectors foreground authenticity and experiential immediacy.

The integration of results shows that the documentary's meaning-making depends on a complementary multimodal relationship. Language delivers propositional content and temporal continuity, while visual and auditory resources enhance vividness, emotional resonance, and narrative coherence. High-modality visuals (saturated color, sharp focus, dynamic framing) work together with authoritative prosody to construct a credible, optimistic image of China's development. Dialogues and interactive framing further invite empathy and identification, transforming factual reportage into an engaging human-centered narrative.

Overall, the synergy between linguistic, visual, and auditory modes achieves both informational and affective persuasion: the documentary not only informs but also evokes, allowing audiences to perceive China's transformation as both visible and relatable. This synthesis demonstrates how quantitative patterns of modal distribution and qualitative semiotic analysis jointly illuminate the documentary's discourse strategy-to render development tangible through multimodal realism and interpersonal proximity.

Research Limitations

Although the research has made some findings, some limitations still exist in the paper. Firstly, there is insufficient research data concerning dynamic discourse. The analysis in the paper only focuses on a single episode from the entire documentary, lacking comparison with other videos. Secondly, the chosen corpus for this study appears relatively brief,

comprising less than 10 minutes. Thirdly, the investigation into the relationship between multi-modal modes lacks thoroughness and depth.

Hopefully, future research could make more progress based on this study. First, future researchers could broaden the scope of analysis by examining a diverse range of documentaries, films or short videos. Comparing the multimodal discourse features across different works would provide a more comprehensive understanding of their impact on information dissemination and audience perception.

Second, future researchers can delve deeper into the relationships between different modalities. This includes conducting detailed analyses of the interactions between visual, auditory, and linguistic modalities, and exploring how they collectively shape audience understanding and experience.

In summary, future researchers could expand the scope of analysis to include diverse media forms, conduct an in-depth exploration of multimodal relationships, which offers a promising avenue for exploration.

Acknowledgments: The figures 3-29 used in this paper are from documentary Discover China: Behind the Rapid Growth released by China Daily. The use of these images adheres to the Fair Use Agreements in Copyright Law.

Declarations

Funding statement: The author did not obtain any funding for this research.

Data availability: All the data are included in the content of the paper.

Competing interest statement: The author reported no conflict of interest.

Additional information: No additional information is available for this paper.

References:

1. Baldry, A., & Thibault, P. J. (2006). Multimodal transcription and text analysis: A multimodal toolkit and coursebook with associated on-line course. Equinox.
2. Barthes, R. (1977). Image, music, text (S. Heath, Trans.). Fontana Press.
3. Bateman, J. (2008). Multimodality and genre: A foundation for the systematic analysis of multimodal documents. Palgrave Macmillan. https://doi.org/10.1007/978-0-230-58232-3
4. Borah, A., Abdullah, H. M., Wei, K., & Huang, R. (2025, July). CliME: Evaluating multimodal climate discourse on social media and

the Climate Alignment Quotient (CAQ). In K. Atwell, L. Biester, A. Borah, D. Dementieva, O. Ignat, N. Kotonya, Z. Liu, R. Wan, S. Wilson, & J. Zhao (Eds.), *Proceedings of the Fourth Workshop on NLP for Positive Impact (NLP4PI)* (pp. 43–61). Vienna, Austria: Association for Computational Linguistics.

5. Duan, D. (2019). Multimodal discourse analysis of the documentary "Amazing China" [Master's thesis, Xi'an International Studies University].
6. Gu, Y. (2015). Multimodal sensory systems and language studies. Contemporary Linguistics, 17(4), 448-469.
7. Halliday, M. A. K. (1985). An introduction to functional grammar. Edward Arnold.
8. Hu, Z. (2007). Multimodality in social semiotic research. Language Teaching and Linguistic Studies, 1, 1-10.
9. Jewitt, C., Bezemer, J., & O'Halloran, K. (2016). *Introducing Multimodality.* Routledge.
10. Kress, G., & van Leeuwen, T. (1996). The grammar of visual design. Routledge.
11. Kress, G., & van Leeuwen, T. (2002). Colour as a semiotic mode: Notes for a grammar of colour. Visual Communication, 1(3), 343-368. https://doi.org/10.1177/147035720200100306
12. Lan, J. (2020). Discourse analysis of multimodal dynamic videos based on ELAN software: A case study. Journal of Jiujiang University (Social Sciences Edition), 39(3), 90-97.
13. Li, Y., & Feng, D. (2017). PPT courseware design and linguistic knowledge construction: A multiliteracies pedagogy perspective. e-Education Research, 38(5), 95-100.
14. Li, Z. (2003). A social semiotic approach to multimodal discourse. Foreign Language Research, 5, 1-8.
15. Li, Z. (2023). Multimodal discourse analysis of Lanzhou city image promotional videos from the perspective of visual grammar [Master's thesis, Lanzhou University of Technology].
16. Liang, Q. (2018). Multimodal discourse analysis of the BBC documentary "Chinese New Year: The Biggest Celebration on Earth" [Master's thesis, Xi'an International Studies University].
17. Liu, Y., & Zhang, H. (2018). Multimodal discourse analysis of political documentaries in shaping national image. Modern Communication (Journal of Communication University of China), 40(9), 118-122.
18. O'Halloran, K. (2004). Multimodal discourse analysis: Systemic-functional perspectives. Continuum.

19. Tang, M. (2016). Multimodal discourse analysis of the BBC documentary “Wild China” [Master's thesis, Shandong Normal University].
20. Wang, L., & Wen, Y. (2008). Multimodal analysis methods in applied linguistics research. Computer-Assisted Foreign Language Education, 5, 8-12.
21. Wang, R.-X., & Taabaldiev, K. (2025). *Exploring the dynamics of multimodal discourse in social media news: Addressing research gaps in interaction, affordances, and temporal patterns*. Forum for Linguistic
Studies, 7(5), 470- 482. https://doi.org/10.30564/fls.v7i5.8542
22. Wei, Q. (2009). Multimodality and multimodal discourse research in visual environments. Science Press.
23. Zhao, X., & Feng, D. (2017). Multimodal metaphor and metonymy in the construction of China's image: A case study of The Economist's political cartoons. Journal of Xi'an International Studies University, 25(2), 31-36.
24. Zhang, D. (2009). Exploring a comprehensive theoretical framework for multimodal discourse analysis. Foreign Languages in China, 6(1), 24-30.
25. Zhang, D., & Yuan, Y. (2011). A study on modal coordination in dynamic multimodal discourse: A case study of TV weather forecast multimodal texts. Shandong Foreign Language Teaching Journal, 32(5), 9-16.
26. Zhang, Z., & Yang, H. (2021). Telling sports stories well: A multimodal discourse analysis of the TV documentary The Chinese Women's Volleyball Team. Journal of Sports and Science, 42(3), 82-88. https://doi.org/10.13598/j.issn1004-4590.2021.03.013
27. Zhu, Y. (2007). Theoretical foundations and research methods of multimodal discourse analysis. Foreign Language Research, 5, 82-86.