

Original Paper

Multimodal Mechanisms in Urban Image Construction: An Audiovisual Analysis of *Gather in Shenzhen*

Liheng Hu¹

¹ Guangzhou College of Commerce, Guangzhou, Guangdong, China

Received: August 22, 2025
2025 doi:10.22158/eltls.v7n6p1

Accepted: October 21, 2025

Online Published: November 12,
2025 URL: <http://dx.doi.org/10.22158/eltls.v7n6p1>

Abstract

*This study applies the integrated theoretical frameworks of visual and auditory modalities derived from multimodal discourse analysis (MDA) to systematically investigate the mechanisms of audiovisual synergy in the official promotional film *Gather in Shenzhen*, produced for the 2026 APEC Economic Leaders' Meeting. It aims to demonstrate how the interaction between visual and auditory elements constructs a cohesive and multidimensional urban narrative. The analysis shows that the film strategically employs visual and auditory resources to achieve strong thematic consistency across six core themes: vibrant and innovative, comfortable and convenient, green and beautiful, safe and resilient, virtuous and culturally advanced, and smart and efficient. The findings further underscore the pivotal role of four cross-modal synergistic mechanisms, including reinforcement, complementarity, contrast, and juxtaposition, in maintaining discursive cohesion and enhancing emotional resonance. Overall, the integration of visual and auditory modalities effectively presents Shenzhen as a dynamic, welcoming, and technologically advanced global city, thereby strengthening the communicative power of urban promotional media and offering useful insights into the construction of contemporary national and urban images.*

Keywords

multimodal discourse analysis, audiovisual synergy, urban narrative, urban image

1. Introduction

In the contemporary context of deepening globalization and digitalization, the construction and dissemination of national and urban images have become essential strategies for enhancing international competitiveness, strengthening regional soft power, and promoting intercultural communication. Mega-events, such as the Olympic Games, World Expos, and high-level summits including APEC and the G20, serve not only as significant platforms for global governance and

multilateral diplomacy but are also widely recognized as strategic sites for host countries to showcase their economic achievements, cultural characteristics, and core values (Roche, 2002; Zhang & Zhao, 2009). Within this process, official promotional documentaries, as one of the most comprehensive and influential forms of media communication, have transcended their informational function to become crucial discursive practices for the construction of political and cultural meanings. These films carry the profound responsibility of self-presentation and image-building, thereby contributing significantly to the formation of a city's identity and its symbolic representation in both domestic and international contexts (Fairclough, 1995; Machin, 2013).

The forthcoming 33rd APEC Economic Leaders' Meeting, scheduled to be held in Shenzhen, China in 2026, represents a salient opportunity for China to demonstrate its capacity for global governance and its image as a responsible major power. Simultaneously, this event provides Shenzhen, a city renowned for its technological innovation, ecological awareness, and openness to global engagement, with a platform to present its urban spirit and developmental philosophy to the international community. In this context, the release of the official preparatory promotional film, *Gather in Shenzhen*, assumes considerable symbolic and communicative significance, serving as an instrument of international dissemination and public diplomacy.

Through the integrated deployment of visual and auditory modalities, *Gather in Shenzhen* systematically constructs a multi-dimensional and layered urban image of Shenzhen. The film conveys six core thematic dimensions, portraying the city as a *vibrant and innovative* urban center, a *comfortable and convenient* place to live, a *green and beautiful* ecological environment, a *safe and resilient* urban space, a *virtuous and culturally advanced* center of civilization, as well as a *smart and efficient* city epitomizing technological rationality. Each thematic dimension is carefully encoded through coordinated visual-auditory semiotic resources, reflecting both the functional and affective goals of urban image promotion.

This study focuses on *Gather in Shenzhen* and employs the theoretical framework of Multimodal Discourse Analysis (MDA) to systematically explicate how the film achieves a multi-layered, cross-modal construction of Shenzhen's urban image. Specifically, the research examines the interrelationships and meaning-making mechanisms among visual elements, such as imagery, color, and composition, and auditory elements, including music, sound effects, and narration. By adopting this approach, the study deepens the understanding of promotional films as instruments for urban image communication, revealing their expressive strategies and cultural implications within contemporary international communication (Machin, 2013; Bateman, 2014). The findings aim to contribute both to the theoretical development of multimodal discourse analysis and to practical strategies for effective urban image dissemination.

2. Theoretical Foundation and Analytical Framework

In contemporary communication studies, the production of meaning in media texts is understood as a dynamic and integrative process that relies on the interaction of multiple semiotic modalities, including visual, verbal, and auditory resources (Kress & Van Leeuwen, 2001; Jewitt, 2009). Within this perspective, Multimodal Discourse Analysis (MDA) provides a comprehensive theoretical framework for investigating how different semiotic systems collectively contribute to meaning-making. The theory posits that analyzing a single mode in isolation is insufficient for understanding the complex mechanisms through which overall meaning is produced and interpreted. Instead, research should focus on how various modalities are selected, organized, and integrated within specific socio-cultural contexts, thereby generating multi-layered and contextually embedded meanings (O'Halloran, 2011; Ledin & Machin, 2020)

2.1 Visual Modality

Building upon the framework of Systemic Functional Linguistics (SFL), as further elaborated by Halliday and Matthiessen (2013) Kress and Van Leeuwen (1996) proposed the model of Visual Grammar (VG), which extends the three metafunctions of SFL, namely the ideational, interpersonal, and textual functions, into the visual domain. This model posits that images convey meaning through three interrelated dimensions: the representational, interactive, and compositional meanings.

2.1.1 Representational Meaning

The initial analytical focus is directed towards *representational meaning*, which examines how visual texts construct and depict reality by representing the relationships among participants, actions, and settings within a visual frame. Within this framework, visual representations are fundamentally categorized into two principal modes: narrative representation and conceptual representation. Narrative representation systematically employs elements such as vectors, participants, and processes to articulate dynamic event structures, a characteristic particularly salient in the visual narratives of promotional films (O'Halloran, 2011). Specifically, the directed movement of characters, the trajectory of their gaze, transitions in spatial arrangements, and the movement paths of objects collectively function as core components that propel the temporal narrative forward. In contrast, conceptual representation emphasizes the depiction of static, categorical relations, including identity, attribution, or symbolic associations. Through deliberate compositional strategies, the selection of specific postures, or the deployment of emblematic imagery, it conveys abstract notions such as national identity, cultural values, or brand ideologies (Jewitt, 2009; Van Leeuwen, 2008). The sophisticated interplay between narrative and conceptual modes thus enables promotional films to present factual content in a compelling and vivid manner while simultaneously conveying underlying ideological and persuasive connotations.

2.1.2 Interactive Meaning

The second dimension of analysis focuses on *interactive meaning*, which examines the social relationships and emotional connections established between visual texts and their viewers. In the

context of promotional films, this dimension is primarily realized through specific visual parameters, including gaze, camera angle, and social distance (Kress & Van Leeuwen, 1996). The deployment of a direct gaze by on-screen participants systematically establishes interpersonal engagement and connection, typically eliciting affective responses such as trust, invitation, or empathy. In contrast, an indirect or averted gaze frames the viewer as an external onlooker, thereby creating a sense of separation or reduced engagement with the scene (Jewitt & Oyama, 2001). The strategic selection of camera angle further conveys symbolic significance. For example, high-angle shots frequently communicate authority, control, or superiority, rendering them appropriate for emphasizing institutional or brand power, whereas low-angle shots express admiration, reverence, or grandeur, and are often utilized in representations of national identity or cultural heritage (Machin, 2013). Additionally, social distance, manifested through close-ups, medium shots, or long shots, regulates the viewer's perceived psychological proximity to the subject, with close-ups enhancing intimacy and emotional involvement and long shots emphasizing spatial context and overall atmosphere (Bateman, 2014).

2.1.3 Compositional Meaning

The final analytical category, *compositional meaning*, addresses the spatial organization and hierarchical arrangement of visual elements within the frame, a process that ultimately determines both the logical flow and informational emphasis conveyed by the image (Kress & Van Leeuwen, 1996). The central components of this dimension include information value, salience, and framing. Information value pertains to the strategic placement and sequencing of elements within the visual field. For example, elements positioned centrally or in concluding areas often carry core or conclusive messages (O'Halloran, 2008). Salience is systematically achieved through the manipulation of visual parameters such as brightness, color contrast, motion, and depth of focus, which function to guide and direct viewers' attention (Jewitt, 2009). Framing refers to the techniques used to connect or separate distinct scenes through editing decisions, shot transitions, and changes in setting, thereby constructing a coherent and structured visual discourse (Bateman & Schmidt, 2013). In the context of promotional films, compositional meaning not only influences the overall narrative pacing and emotional tone but also reflects the ideological orientation and communicative intent of the producers (Van Leeuwen, 2008).

The overarching theoretical and methodological utility of Visual Grammar (VG) lies in its ability to provide a systematic and robust framework for analyzing semiotic selection, visual narrative structure, and emotional expression within dynamic multimodal texts, such as promotional films (Kress & Van Leeuwen, 1996; Jewitt, 2009). This framework posits that the visual mode constitutes a distinct yet analogous social semiotic system parallel to language, capable of realizing meaning through the deliberate selection, configuration, and combination of semiotic resources within specific sociocultural contexts (Kress & Van Leeuwen, 1996). The principal contribution of VG is its systematic elucidation of the underlying mechanisms of visual meaning-making, enabling scholars to interpret the structural

organization, communicative intent, and social effects of visual texts from both semiotic and sociocultural perspectives (Jewitt & Oyama, 2001; O'Halloran, 2008). For the analysis of multimodal dynamic texts, VG serves as an empirically grounded and operationally flexible tool, encouraging researchers to move beyond surface-level image description and to examine how visual meaning is actively constructed through the interaction of semiotic resources and social context. This approach facilitates a nuanced understanding of how visual elements function in conveying information, expressing emotion, and articulating ideological and cultural values (Machin, 2013).

2.2 Auditory Modality

Van Leeuwen (1999) conceptualizes the auditory modality as a distinct semiotic mode with its own meaning-making potential. Although sound differs from both language and image in its formal and perceptual characteristics, which complicates a direct mapping of SFL's three metafunctions onto auditory structures, it nevertheless contributes systematically to meaning construction (O'Halloran, 2008). In the analysis of multimodal dynamic texts, such as promotional films, treating sound as an independent semiotic resource parallel to the visual mode allows for a nuanced understanding of its functions in narrative development, emotional engagement, and the communication of ideological and cultural values (Chion, 2019). Van Leeuwen (1999) further identifies six interrelated analytical dimensions of the auditory modality: *sonic interaction, perspective, rhythm, voice quality together with timbre, melody, and modality*.

Sonic Interaction examines the ways in which sound engages with the audience, the represented scene, and other semiotic modes. This dimension specifically considers whether the sound source is presented on-screen or off-screen, whether it establishes a reciprocal relationship with on-screen participants, and how voice-over, narration, and ambient sound collectively contribute to the construction of the auditory discourse structure (Van Leeuwen, 1999). In the context of promotional films, direct human voices or interview segments that address the audience typically generate a strong sense of engagement and persuasive immediacy, whereas environmental sounds and sound effects are employed to enhance realism and convey emotional significance (Chion, 2019).

Perspective or spatiality refers to the mechanisms through which auditory distance and the spatial positioning of sound sources are conveyed. Analogous to high and low camera angles or long and close shots in visual grammar, auditory perspective is determined by factors such as recording distance, virtual spatial placement within the sound field, and the application of reverberation during post-production. Close-miked voices typically communicate intimacy and direct address, whereas distant or reverberant soundscapes contribute to a perception of spatial expansiveness or reinforce an objective narrative tone (Van Leeuwen, 1999). In promotional films, producers may strategically employ close-range vocal delivery to enhance psychological proximity and facilitate audience identification, or utilize distant ambient sounds to evoke grandeur and convey institutional authority.

Rhythm refers to the temporal organization of the auditory modality, encompassing structural elements such as phrasing, beat patterns, and variations in tempo, all of which play a critical role in shaping

narrative pacing and emotional progression. Rhythm integrates both verbal components, including speech rate and the strategic use of pauses, and nonverbal sound structures, such as musical or percussive beats. The deliberate modulation of rhythmic parameters directly influences the tension and affective trajectory of the promotional narrative: faster tempos typically generate excitement and a sense of mobilization, whereas slower tempos evoke solemnity or encourage emotional reflection (Van Leeuwen, 1999). In addition, the degree of synchrony, or the deliberate use of asynchrony, between visual editing and the accompanying soundtrack serves as a crucial analytical cue for interpreting the complex relationship between image and sound (Chion, 2019).

Voice quality and timbre constitute the dimension that examines how the specific texture and acoustic properties of sound contribute to the nuanced expression of identity, emotion, and attitude. Vocal characteristics, including tone depth, vibrato, and breath dynamics, together with the emotional coloring of speech and the timbral selection of background music and sound effects, collectively function to signal particular social identities, discrete emotional states, or specific cultural styles (Van Leeuwen, 1999; Kassabian, 2002). In promotional films, techniques such as low-frequency enhancement to convey solemnity, or the deliberate use of culturally meaningful timbres in musical accompaniment, are frequently employed to foster audience identification and reinforce cultural affiliation.

Melody examines the mechanisms through which musical features contribute to meaning-making, focusing on characteristics such as pitch contour, harmonic mode, and underlying harmonic structure. These elements function as emotional cues, narrative markers, or leitmotifs, with the latter defined as musical motifs that establish thematic connections with specific narrative ideas or characters across multiple scenes. Empirical studies indicate that the strategic deployment of recurring melodic motifs can serve as cognitive anchors, effectively reinforcing brand memory and enhancing narrative cohesion throughout the filmic text (Kassabian, 2002).

Modality refers to the perceived level of reality or authenticity of sound, indicating the extent to which an auditory text is experienced by the listener as either an accurate representation of a real-world recording or as a deliberately constructed artistic representation. This dimension considers whether the sound is presented as a live recording, preserving ambient noise, irregularities, or performance imperfections, or as a heavily produced text, characterized by extensive editing, manipulation, or dramatic embellishment. Sounds exhibiting high modality enhance realism and a sense of immediacy, whereas low-modality sounds, such as synthesized effects or carefully produced musical scores, primarily function to guide emotional response and achieve aesthetic expression (Van Leeuwen, 1999; O'Halloran, 2008). In promotional film production, creators frequently make strategic decisions regarding the balance between acoustic authenticity and expressive impact in order to achieve persuasive aims and reinforce brand or institutional image.

It is critically important to recognize that the six dimensions of auditory modality are interdependent rather than functioning in isolation, working synergistically to generate complex meaning. For example,

the deliberate combination of a close auditory perspective, warm timbral quality, and slow rhythmic pacing can collectively construct an intimate and trustworthy narrative persona for the audience. In contrast, the interaction between a grand orchestral melody and distant reverberation may evoke a majestic and authoritative public image (Chion, 2019; Van Leeuwen, 1999; O'Halloran, 2008). Consequently, integrating Van Leeuwen's auditory modality framework with established approaches in multimodal discourse analysis and social semiotics provides a systematic and operational methodology for examining how sound interacts with visual elements to shape meaning, guide emotional response, and fulfill the communicative and persuasive objectives of the text.

Building on the theoretical frameworks of visual and auditory modalities, the following analysis examines the synergistic interaction of image and sound in shaping the urban narrative of Shenzhen. It focuses on how visual and auditory resources are orchestrated to construct persuasive meaning, guide audience perception and emotional engagement, and convey the city's socio-cultural, economic, and ideological identity, thereby providing a foundation for a detailed multimodal analysis of *Gather in Shenzhen*.

3. Multimodal Analysis of *Gather in Shenzhen*: Cross-Modal Construction of Urban Image

This study focuses on the promotional film *Discover Shenzhen: Host City of the 2026 APEC Economic Leaders' Meeting*, released via CGTN's official media platforms in November 2025 (hereafter referred to as *Gather in Shenzhen*). By integrating the analytical frameworks of Visual Grammar and Auditory Grammar, the analysis examines the synergistic mechanisms and semiotic strategies through which visual, auditory, and linguistic resources collaboratively construct and communicate Shenzhen's contemporary urban image. This cross-modal approach highlights how the film strategically positions the city as a dynamic, modern, and globally connected urban space in anticipation of the 2026 APEC Economic Leaders' Meeting.

3.1 Visual Modality Analysis

3.1.1 Representational Meaning: the Interweaving of Dynamic Vectors and Conceptual Symbols

The visual representational strategies deployed in *Gather in Shenzhen* are primarily realized through the systematic construction of two complementary dimensions: dynamic kinetic processes and abstract symbolic values, which correspond predominantly to the analytical categories of narrative representation and conceptual representation, respectively. These dual strategies work in tandem to convey both the tangible development of the city and its underlying ethos, providing a multidimensional portrayal of Shenzhen's urban identity.

Narrative representation centrally emphasizes the depiction of the city's ongoing and accelerated development through visible actions and identifiable processes. The film frequently employs pronounced vector structures formed by high-speed motion, including speeding trains, incessant traffic flows, sequences of arriving and departing flights, and busy port operations [00:34–00:42]. These visually dynamic elements, containing clear directional cues, not only generate a vivid sense of

on-screen motion but also symbolically convey Shenzhen's strong developmental momentum and operational efficiency. This narrative strategy reinforces the city's positioning as a regional and global hub marked by dense economic activity and modern infrastructure. In segments dedicated to the theme of *vibrant and innovative* [01:26–01:50], frequent montage editing and rapid shot transitions create a visual rhythm that mirrors accelerated information flow and technological iteration, further consolidating Shenzhen's narrative as an "innovation engine." Narrative representation thus foregrounds temporality, motion, and transformation, communicating the city's dynamism and forward-looking character.

Conceptual representation, in contrast, functions to communicate Shenzhen's spiritual essence, core values, and development philosophy through potent symbolic imagery. The film extensively utilizes strategically framed close-ups and slow-motion shots to endow key visual elements with meanings that transcend their physical forms. For example, the sculpture of the Pioneer Ox [01:01], filmed from a low-angle perspective, visually emphasizes its mass and inherent strength, symbolically highlighting the city's exploratory spirit and resilient, determined character. Similarly, the towering skyscrapers and formations of flying drones [00:21; 03:51], with their strong vertical lines and ascending spatial composition, convey the conceptual image of Shenzhen as a modern, efficient, and intelligent city. The sophisticated interplay between narrative and conceptual representations enables the film to integrate depictions of tangible urban activity with abstract symbolic values, constructing a cohesive and persuasive multidimensional urban identity tailored for the APEC 2026 audience.

3.1.2 Interactive Meaning: Dynamic Regulation between Grand Narratives and Equitable Engagement

At the level of Interactive Meaning, *Gather in Shenzhen* skillfully balances the projection of grand, authoritative urban narratives with the need to cultivate human-centered proximity and equitable engagement. This balance is achieved through the deliberate regulation of visual parameters, including social distance, camera perspective, and gaze relationships, thereby effectively supporting the construction of the *comfortable and convenient* as well as the *safe and resilient* urban themes.

On a macro scale, the film extensively employs high-angle aerial shots and extreme long shots [00:15, 00:22, 00:41, 02:33] to convey Shenzhen's vast spatial order and structural coherence. This panoramic visual strategy not only imparts a sense of comprehensive control and stability but also reinforces the city's image as a safe and resilient metropolis by emphasizing its structural order and comprehensive oversight. The distant, non-personal perspectives afford viewers a slightly elevated, panoramic view, allowing them to apprehend the city's architectural and infrastructural organization while simultaneously framing Shenzhen as a model modern urban space.

To modulate the authoritative quality of these macro perspectives, the film strategically intersperses medium and close-up shots during key narrative sequences, particularly in scenes depicting citizen leisure activities, cross-cultural interactions, and daily life [01:55, 03:08]. The adoption of an eye-level perspective and reduced social distance establishes a more intimate and equitable visual relationship, transforming abstract urban concepts such as livability, inclusiveness, and cultural vibrancy into

concrete, relatable experiences. Notably, medium shots within technology and innovation segments [01:38] foreground human agency, ensuring that Shenzhen's technological and innovative identity remains approachable rather than impersonal.

Gaze relationships further enhance these interactive shifts. In professional work and efficiency-focused scenes [01:32–01:34], subjects do not look directly at the camera, which emphasizes their professionalism and operational competence while positioning the viewer as an observer of the city's activity. In contrast, the emotionally resonant concluding sequence [04:17] shows citizens facing the camera, smiling and waving, creating a sense of direct engagement and invitation. This deliberate change in visual address transforms Shenzhen from a city presented as a distant, observational subject to one that actively interacts with the viewer, significantly enhancing its affective appeal through emotional connection and interpersonal resonance.

Through the strategic orchestration of spatial perspective, social distance, and gaze, *Gather in Shenzhen* dynamically negotiates between grand, authoritative narratives and human-scale engagement. This nuanced regulation not only reinforces the city's overarching narrative of modernity, safety, and efficiency but also ensures that Shenzhen is presented as an inclusive, approachable, and emotionally resonant urban environment, aligning seamlessly with the promotional objectives for the forthcoming 2026 APEC meeting.

3.1.3 Compositional Meaning: Constructing Color Semiotics and Orderly Urban Frameworks

At the level of Compositional Meaning, *Gather in Shenzhen* establishes a visually harmonious and highly organized structure that is both temporally sequenced and spatially coherent through the coordinated deployment of salience, color semiotics, and information value. This compositional strategy systematically reinforces the city's overarching thematic concept of high-quality development, conveying complex urban narratives.

Color, as a central visual semiotic resource, serves the dual function of thematic categorization and emotional evocation. The film employs a carefully curated palette dominated by three primary color schemes, namely technological blue, ecological green, and festive red, each strategically aligned with specific thematic segments.

Technological Blue: This cool visual tone dominates the segments focusing on the *vibrant and innovative* as well as the *smart and efficient* city themes. [01:26–01:50; 03:10–03:27]. Blue hues emphasize rationality, precision, and a technological aesthetic, resonating with Shenzhen's characteristics of innovative vitality and operational efficiency. The deliberate use of sharp lighting and high contrast further enhances the visual salience of technological elements, positioning innovation as a central driver of the city's narrative.

Ecological Green: Saturated natural greens and oceanic blues permeate the scenes representing *comfortable and convenient* as well as *green and beautiful* cities [01:52–02:41]. These colors jointly construct an environmentally conscious and visually refreshing urban landscape, conveying sustainable development and the integration of urban life with natural surroundings. The soft focus, balanced

lighting, and harmonious color gradation in these segments emphasize livability and visual comfort, complementing the city's ecological identity.

Festive Red: Appearing prominently in the *virtuous and culturally advanced* segments [03:26–03:39], red functions as a cultural signifier. Dynamic cultural elements, such as the vibrant costumes of lion dances and the marked crews in dragon boat racing, leverage red to strengthen regional cultural identity, evoke collective pride, and intensify emotional engagement with the city's cultural heritage.

In terms of information value and compositional hierarchy, the film achieves a visual equilibrium between economic development and livability through spatial organization and directional structuring. High-angle aerial shots frequently adopt a top-down perspective, juxtaposing towering skyscrapers with expansive urban green spaces. This symmetrical spatial arrangement visually reinforces Shenzhen's dual identity as a *safe and resilient* city, as well as a *comfortable and convenient* one, materially embodying the multidimensional concept of high-quality urban development.

Additionally, the film foregrounds horizontal connectivity through deliberate framing of transport infrastructures, including highways, bridges, and flight paths [00:30–00:41]. These extended visual axes emphasize Shenzhen's accessibility, openness, and integration as a regional and global hub. The interplay of vertical and horizontal compositional strategies not only organizes the visual field effectively but also communicates the city's simultaneous dynamism, structural order, and commitment to sustainable, high-quality urban growth.

3.2 Auditory Modality Analysis

The auditory modality in *Gather in Shenzhen* functions as a crucial semiotic resource for evoking emotion, shaping discursive coherence, and constructing a multifaceted urban image. The film achieves its narrative logic through the systematic, multidimensional coordination of music, sound effects, and narration, reflecting the integrated roles of representational, interactive, and compositional functions.

3.2.1 Auditory Representation and Thematic Identity

From the representational perspective, auditory elements, including sound effects, melody, and timbre, collectively construct and anchor Shenzhen's diverse urban identity. Grand symphonic orchestration underpins the main thematic sections (00:17–01:20), producing a solemn and uplifting epic tone that metaphorically conveys the city's strategic significance and historical contributions within national development initiatives. This sonic scaffolding reinforces Shenzhen's positioning as a *vibrant and innovative* metropolis, as well as a *safe and resilient* one.

Cultural Fusion: In the opening and cultural exhibition segments [00:01–00:08; 03:26–03:39], traditional Chinese musical motifs are deliberately incorporated, creating a harmonious auditory blend of tradition and modernity. These elements resonate with the *virtuous and culturally advanced* theme, projecting robust cultural confidence and an international outlook.

Technological Modality: Sections highlighting innovation and smart technology [01:26–01:50; 03:10–03:20] employ synthesizer effects and data-flow sounds, constructing a high-modality,

technologically mediated auditory landscape. This sonic strategy reinforces the *smart* and efficient identity, emphasizing Shenzhen's cutting-edge technological competence and futuristic urban aesthetic. Ecological Authenticity: In the segment depicting ecological and sustainable development [02:10–02:41], natural sounds, such as birdsong and ocean waves, create a high-modality, environmentally authentic soundscape. These auditory cues align closely with the visual imagery, fostering cross-modal complementarity and reinforcing Shenzhen's ecological consciousness as well as its image as a *green and beautiful* city.

3.2.2 Auditory Interaction and Emotional Modulation

At the interactive level, auditory elements guide psychological and emotional engagement through the modulation of rhythm, melody, and voice quality.

Excitement and Momentum: Rapid crescendos and accelerated rhythmic patterns in segments portraying urban development and technological advancement [01:03–01:15] generate dynamic auditory motion, evoking excitement, pride, and mobilization. This reinforces Shenzhen's identity as a *vibrant and innovative* city.

Tranquility and Balance: Conversely, in the ecological segment [02:10–02:41], slower rhythms, reduced volume, and soft string timbres establish a serene and contemplative atmosphere. This sonic contrast balances the preceding high-energy sequences, effectively conveying the composed, livable qualities of a *comfortable and convenient* city.

Authoritative Persona: The narration, delivered with calm timbre, balanced intonation, and clear articulation, constructs an auditory persona that is authoritative, professional, and trustworthy. This aligns with Shenzhen's portrayal as a *safe and resilient* urban space and supports the film's international communication strategy, projecting institutional credibility and reliability.

3.2.3 Auditory Composition and Discursive Coherence

At the compositional level, the auditory modality ensures overall discursive coherence through the coordinated interplay of foreground and background layers and the continuity of thematic music.

Narration consistently occupies the auditory foreground, ensuring clarity and accurate information delivery. Music and sound effects provide the structural and emotional background, maintaining affective continuity. In segments employing rapid montage and discontinuous visual cuts [00:30–00:44], continuous thematic music acts as an auditory discursive adhesive, integrating fragmented visuals into a coherent semantic and emotional experience. In particular, sections featuring innovation and smart technology benefit from this auditory cohesion [01:25–01:50], which mitigates the potential fragmentation caused by fast-paced editing, shaping Shenzhen's complex urban narrative into an integrated and compelling audiovisual experience for a global audience.

3.3 Mechanisms of Audiovisual Modality Synergy

Gather in Shenzhen constructs an integrated and multidimensional urban image that harmonizes technological rationality with humanistic warmth through the systematic and synergistic interplay of visual and auditory modalities. The film's multimodal sophistication is reflected not only in the

co-occurrence and temporal juxtaposition of audiovisual elements but, more significantly, in intertextual resonance at the level of meaning and the functional complementarity between the two modes. This audiovisual synergy can be analytically articulated through four interrelated mechanisms: reinforcement and resonance, complementarity and bridging, contrast and balance, as well as juxtaposition and integration, each collectively contributing to the film's overarching communicative and persuasive objectives.

The reinforcement and resonance mechanism is characterized by the synchronous convergence of auditory and visual elements at pivotal emotional or rhythmic junctures, amplifying both semantic and affective impact. For instance, in the concluding segment [04:17–04:20], the collective demanded gaze of citizens toward the camera visually coincides with the climactic swell of the orchestral melody. This precise audiovisual synchrony establishes a decisive emotional apex and rhythmic closure. Such co-occurrence not only conveys the city's vibrant character but also transforms the presentation of information into direct emotional resonance, reinforcing Shenzhen's confident, approachable, and humanistic image.

The complementarity and bridging mechanism operates to maintain narrative coherence and regulate cross-modal information rhythm. In the *vibrant and innovative* segment [01:26–01:50], rapidly sequenced technological visuals appear highly fragmented. Yet, the continuous electronic rhythms and sustained symphonic themes in the auditory layer provide essential discursive cohesion and emotional guidance. This complementary function bridges visual discontinuities, enabling the semantic integration of abstract concepts, such as technological innovation, smartness, and efficiency, into a unified, coherent audiovisual message.

The contrast and balance mechanism is employed to articulate dualistic harmonies, such as economic development versus ecological preservation, or speed versus tranquility, through carefully coordinated audiovisual modulation. In the *green and beautiful* segment [02:10–02:41], visuals transition from urban skylines to expansive natural landscapes, while the music shifts from bright brass tones to soft string textures. The combined slowing of rhythm and timbral transformation aligns precisely with visual changes in color, composition, and framing, generating a controlled tensional contrast. This audiovisual interplay effectively communicates Shenzhen's commitment to sustainable urban development, demonstrating a balanced high-quality growth strategy that integrates humanistic, technological, and ecological considerations.

The juxtaposition and integration mechanism is most evident in the *virtuous and culturally advanced* segment [03:26–03:39]. Here, traditional folk practices, including lion dances and dragon boat races, are visually juxtaposed with contemporary urban infrastructure, creating a temporally and spatially layered narrative. Audibly, traditional Chinese melodies are seamlessly integrated with Western symphonic harmonies, achieving a cross-cultural co-construction of tradition and modernity. This multimodal juxtaposition not only foregrounds Shenzhen's identity as a culturally confident and

globally oriented city but also metaphorically reflects its strategic self-positioning as a site of successful cultural symbiosis within a contemporary globalized context.

Through these four mechanisms, *Gather in Shenzhen* exemplifies a sophisticated orchestration of audiovisual resources. The film demonstrates how the strategic synergy of visual and auditory modalities can construct a cohesive, multidimensional urban narrative, evoke emotional resonance, and simultaneously convey Shenzhen's technological vitality, ecological awareness, cultural sophistication, and human-centered values to a global audience.

4. Conclusion

4.1 Key Findings

Grounded in the theoretical framework of Multimodal Discourse Analysis (MDA), this study provides a systematic and in-depth examination of how the promotional film *Gather in Shenzhen*, produced in the lead-up to the 2026 APEC Economic Leaders' Meeting, constructs and conveys the city's multidimensional image to a global audience. The findings indicate that the film successfully establishes a cohesive urban representation through the strategic deployment and sophisticated intermodal integration of visual and auditory resources.

The analysis demonstrates notable semiotic consistency and affective intertextuality across the six core thematic dimensions of the film. *Vibrancy and innovation* are expressed through dynamic visual sequences and rapid musical rhythms, effectively conveying Shenzhen's forward-moving energy and developmental momentum. *Comfort and convenience* are communicated via eye-level perspectives, reduced social distance, gentle melodies, and soft timbres, fostering intimacy and a sense of comfort for the audience. *Greenness and beauty* are articulated through ecological color schemes and authentic natural soundscapes, presenting the city as environmentally conscious and aesthetically harmonious. *Safety and resilience* are reinforced by grand compositional structures, high-angle perspectives, stable timbres, and authoritative narration, symbolizing urban order, infrastructural strength, and institutional reliability. *Virtue and cultural advancement* are conveyed through the integration of traditional symbols with national musical elements, highlighting Shenzhen's historical depth and cultural confidence. Finally, *smartness and efficiency* are signified through techno-blue visuals and high-modality electronic sound effects, embodying the technological rationality of a digitally advanced metropolis.

Crucially, the film leverages cross-modal synergistic mechanisms to integrate these thematic elements into a unified, persuasive urban narrative. Reinforcement and resonance operate at emotional peaks to amplify the intensity of meaning transmission, as exemplified in the concluding segment, where citizens' collective gaze coincides with a climactic musical swell, generating a powerful audiovisual apex. Complementarity and bridging maintain discursive cohesion by offsetting the fragmentation of rapid visual editing with continuous auditory motifs, ensuring semantic and affective integration across modalities. Contrast and balance articulate dualities, such as development versus ecological

preservation and speed versus tranquility, through the coordinated modulation of visual composition, color, rhythm, and timbre, effectively conveying the city's commitment to sustainable and balanced growth. Juxtaposition and integration are employed to construct cultural meaning, merging traditional and contemporary elements across both visual and auditory modalities, particularly in scenes featuring lion dances, dragon boat races, and urban infrastructure, thereby presenting Shenzhen as a city that harmonizes heritage with modernity within a globalized context.

Through these integrated audiovisual strategies, *Gather in Shenzhen* successfully presents the city as a cohesive and multidimensional entity that embodies dynamism and warmth, technological sophistication and human-centered values, as well as tradition and futurity, effectively shaping both perceptual and affective engagement for an international audience.

4.2 Research Limitations and Future Directions

Despite the systematic analysis conducted in this study, several limitations point to important directions for future research. First, the study examines a single promotional video, which limits the richness and generalizability of the findings. Future research could construct a comprehensive corpus, particularly focusing on materials designed to shape national and city images. This would enable broader comparative analyses and provide deeper insights into multimodal representation patterns across diverse contexts. Another limitation is that the current study relies primarily on qualitative methods, which, while effective for in-depth interpretation, could be complemented by quantitative approaches in future work to enhance the robustness of the findings. Additionally, this study concentrates exclusively on the video text itself and does not address audience perspectives or interpretations. Subsequent research could incorporate audience studies to explore how viewers perceive, interpret, and respond to multimodal urban representations, offering a more complete understanding of their communicative effectiveness and impact.

In summation, although this study has certain limitations, its theoretical and practical implications provide useful guidance for future research. It is recommended that subsequent studies further advance this line of inquiry by expanding research corpora, adopting innovative and rigorous methodologies, and incorporating cross-cultural comparisons. Such efforts would contribute to the continued development of multimodal discourse analysis, particularly in the areas of urban image communication, visual culture studies, and national discourse construction.

References

- Bateman, J. (2014). *Text and image: A critical introduction to the visual/verbal divide*. Routledge.
- Bateman, J., & Schmidt, K. H. (2013). *Multimodal film analysis: How films mean*. Routledge.
- Chion, M. (2019). *Audio-vision: sound on screen*. Columbia University Press.
- Fairclough, N. (1995). *Media discourse* (pp. 9-14). London: Edward Arnold.
- Halliday, M. A. K., & Matthiessen, C. M. (2013). *Halliday's introduction to functional grammar*. Routledge.

- Jewitt, C. (Ed.). (2009). *The Routledge handbook of multimodal analysis* (Vol. 1). London: Routledge.
- Jewitt, C., & Oyama, R. (2001). Visual meaning: A social semiotic approach. *The handbook of visual analysis*, 151.
- Kassabian, A. (2002). *Hearing film: Tracking identifications in contemporary Hollywood film music*. Routledge.
- Kress, G., & Van Leeuwen, T. (1996). *Reading images: The grammar of visual design*. Routledge.
- Kress, G., & Van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Arnold.
- Ledin, P., & Machin, D. (2020). *Introduction to multimodal analysis*. Bloomsbury Publishing.
- Machin, D. (2013). What is multimodal critical discourse studies?. *Critical discourse studies*, 10(4), 347-355.
- O'Halloran, K. L. (2011). Multimodal Discourse Analysis. In K. Hyland, & B. Paltridge (Eds.), *The Continuum Companion to Discourse Analysis*.
- O'Halloran, K. L. (2008). Systemic functional-multimodal discourse analysis (SF-MDA): Constructing ideational meaning using language and visual imagery. *Visual communication*, 7(4), 443-475.
- Roche, M. (2002). *Megaevents and modernity: Olympics and expos in the growth of global culture*. Routledge.
- Van Leeuwen, T. (1999). *Speech, music, sound*. Palgrave Macmillan.
- Van Leeuwen, T. (2008). *Discourse and practice: New tools for critical discourse analysis*. Oxford university press.
- Zhang, L., & Zhao, S. X. (2009). City branding and the Olympic effect: A case study of Beijing. *Cities*, 26(5), 245-254.