Appendix 1. Qualitative Research Coding Process

This appendix provides a detailed account of the grounded theory coding process adopted in the first stage of the study, ensuring methodological transparency and traceability. The coding procedure followed the principles of theoretical sampling and constant comparison, and was conducted using MAXQDA 2020. The process ultimately produced a variable framework that served as the empirical foundation for subsequent quantitative analysis.

Open coding constituted the initial step of qualitative analysis, aiming to extract preliminary concepts from the raw interview and textual data. The study analyzed 26 interview transcripts and 31 production tutorial texts line by line, identifying a total of 28 initial concepts. These concepts represent the smallest units of meaning emerging from the data, each closely related to the narrative mechanisms of tourism short videos. They provided the conceptual groundwork for subsequent categorization and theoretical integration.

During the axial coding stage, the initial concepts derived from open coding were systematically clustered and connected. Based on semantic affinity, the 28 concepts were aggregated into nine first-order categories. These categories were then mapped onto two overarching dimensions—narrative content (NC) and narrative discourse (ND)—in accordance with established narrative theory frameworks, thereby establishing the hierarchical structure of the core variables. This step connected the empirical findings with theoretical insights, providing a clear conceptual anchor for the design of quantitative variables in later stages.

Selective coding aimed to integrate all categories into a coherent analytical framework capable of explaining the core phenomenon. Grounded in media ecology theory and complexity theory, the process consolidated NC and ND as the central analytical categories. The resulting framework posits that NC and ND interact dynamically within the viewing context, forming multiple interactional configurations that, through synergistic effects, drive sustained viewing behavior. This core framework not only delineates the complementary relationship between the value transmission of NC and the experiential shaping of ND but also provides a robust theoretical foundation for subsequent quantitative validation and mechanism testing.

Table 1 Coding Process

Core Category	First-Order Category	Initial Concept	Representative Excerpt
Narrative discourse	Perspective	Agent Perspective	◆*Try shooting from different points of view, such as a first-person travel perspective, where the footage may include selfies, scenes of enjoying local cuisine, or appreciating landscapes. The camera must always be ready to capture interesting moments in real time. It's crucial to record them immediately, because what happens right here and now is the most authentic*——XHS011
		Cinematic Perspective	◆*I think the Soochow one is a particularly distinctive short video. Perhaps it's because Suzhou itself is beautiful, but the video uses a style reminiscent of advertisements or films. Some of its footage features slow motion or close-ups—it's quite refined. For example, in a rainy Jiangnan scene, you see a slow-motion shot of a water droplet falling into a lake as a small boat passes by. It really gives you the urge to visit. This kind of cinematic presentation is very appealing to me*——P020
		Viewport Perspective	◆*In terms of shooting style, I personally prefer landscape (horizontal) format. When I was making a video last time, I found that portrait (vertical) format tends to look unbalanced and lacks smoothness. In contrast, landscape framing feels much more natural. So in future shooting, I'll stick to horizontal mode because it produces better visual outcomes*——P005
		Character Perspective	◆*First, I prefer tourism short videos that present unique perspectives. For example, if a blogger visits Zibo and simply says the barbecue is delicious and the locals are nice, that's no different from others' videos. But if the blogger went before Zibo became viral, and returns later to see if the city has changed due to online hype, that's a more compelling angle. Similarly, with Antarctica—if a video just shows glaciers or tough journeys, it's generic. But if the blogger shares cruise experiences or something unexpected like witnessing someone swimming naked, it adds uniqueness. So the key is a distinctive perspective—either in the overall angle or in specific moments of the journey*——P024
	Rhythm	Cutting Rhythm	◆*Equally suitable for fast-paced travel videos, our go-to lazy editing technique - repetitive action cuts stitch together similar motions across different scenes. This works exceptionally well for long journeys incorporating diverse landscapes and terrains, such as road trips or compiling multiple travels into one visually stunning sequence*——XHS002
		Sonic Rhythm	◆*Music remains our most effective emotional catalyst. Most music platforms now categorize by mood, allowing direct filtering to select and download fitting tracks*——XHS016
		Semantic Rhythm	◆*In this fairy-tale world of verdant greenery, as the video approaches its climax, the pacing intensifies with added percussion. Gliding across water amidst emerald hues, synchronized with the rhythmic beats, the runner becomes Dorothy entering Oz. Who says Yangzhou is only for spring? Its summer fairy tale unfolds on screen—*

Core Category	First-Order Category	Initial Concept	Representative Excerpt
			—DY010
	Structure	Unit Style	◆*Normally, I divide the content creation into two or three parts. For example, I used to prefer the purely video-focused format—short, snapshot-like clips. But later, I realized that if it's just one video without captions or supplementary footage from different angles, it feels monotonous and rather dull*——P5
		Continuous Style	◆*I tend to favor a sequential approach because, after all, I'm not here to watch a movie. If you throw in too many disjointed elements, it just feels messy. I want a clear, straightforward flow—something intuitive and easy to follow*——P020
		Discrete Style	◆*When I watch travel short videos, I'm drawn to the highlights. My time is precious, and I don't have the patience to sit through lengthy content*——P006
	Genre	Itinerary Oriented Content	◆*For instance, I have come across short videos featuring 'special forces-style' food tours. I occasionally watch them, but usually only when I have already decided to visit that destination. Such videos effectively provide me with a checklist I can refer to, making them highly purposeful and practical*——P025
		Cultural Interpretation Content	◆*Among various elements, I believe this form of presentation most effectively highlights the local distinctiveness. Take Japan as an example: its unique cultural landscapes are a typical representation. There are also unclassified elements, such as a scene at the Shibuya Crossing, where the background music is carefully mixed—sung by Japanese artists, while the foreground resembles a station announcement like 'the next station is Shibuya.' Immediately following is the bustling scene of pedestrians crossing and flashing billboards. These components together create an atmosphere imbued with Japanese cultural characteristics, making one feel as if standing on the streets of Japan and experiencing its cultural charm firsthand*——P020
		Gastronomic Exploration Content	◆*I really want to visit Quanzhou! I once watched a program about Quanzhou's fish balls and other coastal delicacies from Fujian and Zhejiang, such as Chaoshan beef balls. It made my mouth water, and I was almost ready to set off immediately—honestly, I nearly went right then*——P006
		Experiential Narrative Content	◆*I think the message he conveys is simply about recording his own happiness*——P022
		Emotive Expression Content	◆*But usually, I feel like I bring my own emotions into it. For example, when I post a video, if I'm in a good mood, I tend to choose positive and optimistic content; if I'm in a bad mood, I might select something with a more negative tone*——P012
		Flash Compilation Content	◆*Take Beijing as an example: if one intends to introduce the city, it is necessary to show landmarks such as Tiananmen Square, the Forbidden City, and the Great Wall. If only a few of these are presented, it feels incomplete*——P023
	Drama	Curiosity Order Drama	◆*For the opening, you can use a message exchange or typing chat format to introduce the travel destination*——DY003
		Surprise Order	◆*I once saw a video where, at the end, the creator*

Core Category	First-Order Category	Initial Concept	Representative Excerpt
		Drama	*exaggeratedly talked about how much they ate at a certain place—claiming they were so full they needed an ambulance! The over-the-top humor made the local food seem incredibly memorable and delicious*——P020
	Linguistic Style	Functional Declarative	◆*When exploring a place I've never been, I look for guide-style content—key attractions, must-try foods, and other high-value details. This type of information helps me plan efficiently and grasp the local highlights*——P020
		Affective Narrative	◆*During editing, insert captions like: "At Moon Factory, I spotted red pandas; in Panda Villa No. 1, I met this super cute giant panda—its name is [XXX]." Then, review the video for coherence and flow. Boost engagement with emojis and beat-sync effects, and polish the final cut for maximum impact*——P013
		Imagistic Semiotic	◆*Some creators avoid appearing on camera or speaking. Instead, they rely on music switches to signal transitions, paired with text overlays to set the scene (e.g., "Current vibe: [mood/location]"). This minimalist style can be surprisingly effective*——P025
Narrative Content	Plot Logic	Journey Progression	◆*If the morning weather is good, be sure to capture some aerial shots to establish the environment, then film the vehicle setting off. We spent about fifteen minutes shooting the opening sequence of a travel clip, then made a quick edit to show everyone. For the opening shot—step forward a little, just to set the scene*——DY001
		Scenic Transposition	◆*It was a Christmas decoration at Pavilion, an international mall in downtown Kuala Lumpur, Malaysia. I chose that background music because the scene itself was visually striking and would capture viewers' attention*——P001
		Causal Logic	◆*I prefer more detailed travel guides, like the two-day-one-night itinerary I mentioned for Changsha. Once I decide on a destination, I value content that specifies exactly where to go and what to eat on each day*——P008
	Perceived Authenticity	Genuine Authenticity	◆*I absolutely must visit this place in my lifetime—Tibet. Its unique cultural heritage and breathtaking landscapes are utterly captivating. There's something about Tibet's natural beauty that feels irreplaceable*——P005
		Alterreal Authenticity	◆*First, I applied a crayon-style filter, then layered text with a tilt and drop shadow to enhance the visual impact. The final touch balanced artistry and clarity*——P009
	Content Value Orientation	Hedonic Orientation	◆*I want my videos to feel lively, bright, and full of happiness. Honestly, most of my travel experiences have been incredibly joyful—it's during these trips that I truly feel alive, compared to my usual routine at school. That's why I love documenting these moments. Not just for myself, but to share that infectious energy with others. Maybe they'll feel that same spark of joy too!*——P002
		Utilitarian Orientation	◆*The real utility? Hardcore travel guides—itinerary breakdowns, must-pack items, attraction tips, and even the best photo spots. Pure gold for fellow travelers*——DY014

Appendix 2.Variable Description and Measurement

2.1 Narrative Discourse

2.1.1 Agent Perspective

In video narration, determining who speaks can directly reflect the narrative distance and immersion intensity between creators and audiences. Based on the consensus in narrative linguistics that personal pronouns are the most direct grammatical markers of perspective (Gu & Tse, 2016; Nan et al., 2017), this study conducted automated processing on the audio texts of 18,488 tourism short videos, extracting the occurrence ratios of first-person, second-person, and third-person pronouns, and thereby determining the dominant perspective (Gu & Tse, 2016). When the count of all pronouns is zero, it is recorded as non-personal; if pronouns exist, the one with the highest proportion is taken as the main perspective. For the convenience of subsequent quantitative analysis, the results were coded into a single categorical variable: first person = 0, second person = 1, third person = 2, non-personal = 3. This indicator reveals the differences in narrative perspective in tourism short videos in an objective and replicable manner, providing reliable data support for exploring the impact of agent perspective on user behavior.

2.1.2 Cinematic Perspective

To reveal the narrative differences in tourism short videos regarding who is speaking and whether they appear in the shot, this study coded cinematic perspective into four-level discrete labels: on-camera real-time explanation = 0, third-party voice-over narration = 1, in-frame multi-person dialogue = 2, and no voice/pure scene = 3. The quantification approach is based on a core principle—simultaneously examining the sound source and on-screen presence. In the sound dimension, the paper estimates the speech duration and determines the number of speakers to distinguish between monologue, dialogue, and silence; in the visual dimension, it detects lip movements and facial matching to judge whether the speaker appears in the shot. The cross-analysis of these two dimensions can reliably map to the above four labels without relying on manual annotation. Compared with traditional unimodal methods that only use subtitles or audio tracks, this scheme integrates dual information of the speaking subject and on-screen presence, which aligns with narrative theory and features computational efficiency and robustness. It can stably output cinematic perspective variables directly incorporated into quantitative models for sample sizes at the ten-thousand level.

2.1.3 Viewport Perspective

To quantify the presentation form of short videos on terminals, this study maps screen geometric features into four-level discrete labels: vertical screen (0), horizontal screen (1), square screen (2), with the discrimination based on two stable physical attributes: resolution and projection markers. This classification is consistent with the industry's common standards for vertical/horizontal screen thresholds (Han et al., 2024) and can accurately cover immersive panoramic content. The required information can be directly extracted from video metadata or single-frame dimensions, so the measurement process does not depend on specific screen content, featuring high reproducibility and cross-platform consistency. It can be an independent variable for presentation methods in subsequent quantitative models.

2.1.4 Character Perspective

This study maps the relative distance and orientation between the camera and the subject in

the frame into four-level discrete labels: first perspective (camera position close to the shooter himself, such as hand-held selfies) = 0, second perspective (close-up interviews where the character faces the camera) = 1, third perspective (characters in long shots or side shots, with the audience watching as onlookers) = 2, and no-person perspective (no recognizable characters in the frame) = 3. The judgment logic integrates the area proportion of human faces and bodies in the entire frame, orientation posture, and suspected hand-held features, with priority ranging from close to far, and then to no person. This quantification method replaces traditional semantic interpretation with measurable visual geometric features (Araujo et al., 2020), , which conforms to the concept of subject-object distance in narratology and maintains consistency across scenes and languages. At the same time, it outputs dominant labels and the proportion of frames in each category, providing both discrete and continuous indicators for subsequent models, and can efficiently and robustly depict differences in character perspective on large-scale short video samples.

2.1.5 Rhythm

This study characterizes the speed and intensity of short videos through four complementary indicators: musical beats, cutting density, motion density index, and sound-visual synchronization. Musical beats and cutting density describe the coupling intensity of music and editing. In contrast, motion density index and sound-visual synchronization reflect the dynamic tension between images and content. Intervals normalize these four indicators and then average arithmetically to form a continuous, comprehensive rhythm index. Integrating auditory, visual, and temporal coupling information simultaneously, the continuous comprehensive rhythm index can capture the emotional pulses driven by soundtracks and quantify the rhythmic tension arising from editing and motion. As a single continuous variable, it can be directly incorporated into subsequent quantitative models without further discrete classification, thus avoiding subjective grading and maintaining cross-sample comparability and replicability.

2.1.6 Structure

This study defines the structural form of short videos as the visual and semantic continuity between adjacent narrative segments. It uses the average cosine similarity of segments as a quantitative measure. Based on this indicator, the degree of narrative coherence is divided into unit style (label 0), continuous style (label 1), and discrete style (label 2) from high to low. Measuring the connection intensity of segments through deep visual features can intuitively map the three-part theory of continuity-unit-fragment without relying on manual annotation, providing an objective and reproducible structural form variable for subsequent quantitative models.

2.1.7 Genre

This study regards the genre of tourism short videos as the comprehensive presentation of narrative intentions in text, images, and audio tracks. It classifies samples into six categories based on multimodal clues: flash-compilation content (0), emotive expression content (1), gastronomic exploration content (2), itinerary-oriented content (3), cultural interpretation content (4), and experiential narrative content (5). The model simultaneously reads keywords in subtitles/narration, thematic elements, movement rhythms in images, and beat-emotion features of soundtracks or voices, forms a genre confidence vector through weighted fusion, and takes the one with the highest score as the main label. This strategy replaces subjective scene interpretation with verifiable cross-modal features, which retains the distinction of narratology on differences in purpose-emotion-rhythm and has the dual advantages of large-scale automated processing and

result interpretability.

2.1.8 Drama

This study defines the drama of short videos as the combination of suspense and surprise dimensions presented in the narrative tension of the work. It forms a four-level label through the integration of multimodal signals (text emotional jumps, suspense/surprise keywords, audio-visual mutation intensity): no dramatic structure (0), curiosity-order drama (1), surprise-order drama (2), curiosity + surprise-order drama (3). This quantitative framework matches the drastic fluctuations of the emotional curve with high-frequency trigger words in the field. It introduces objective indicators such as visual quick cuts, color variations, and audio energy transitions. It retains the core focus of narratology on foreshadowing, reversal, and climax. It avoids the subjectivity of manual interpretation, providing a replicable and interpretable measurement benchmark for analyzing narrative tension in large-scale tourism short videos.

2.1.9 Linguistic Style

This study conceptualizes the linguistic style of subtitles in tourism short videos into three categories: Functional Declarative (0), Affective Narrative (1), and Imagistic Semiotic (2). At the sentence level, the model integrates keyword matching and heuristic feature weights to score three dimensions: information guidance, emotional expression, and atmosphere rendering, and outputs the dominant label based on the highest proportion of sentences in the entire video. This dictionary-rule framework maintains high sensitivity to pragmatic expressions such as transportation, prices, and travel guides, while also being capable of identifying internet-based emotional terms and symbolic rhetoric. Its computational logic is lightweight, with each video processed in only a few seconds, enabling stable operation on samples at the ten-thousand level. It combines interpretability and efficiency, providing reliable quantitative support for comparing information services' intensity and emotional mobilization's effectiveness in tourism narratives.

2.2 Narrative Content

2.2.1 Plot Logic

Based on grounded analysis, this study defines plot logic as three coherent dimensions—time, space, and causality—and accordingly constructs a comprehensive indicator, PLI (Plot Logic Index). The temporal dimension measures the richness and sequence of time information in the work to determine whether the narrative progresses along the time axis; the spatial dimension evaluates the coherence of travel routes by combining the recognition of tourism destination names and geodesic distances; the causal dimension depicts the completeness of narrative causality through the density and coverage of logical connectives. The model extracts corresponding features from text and image channels, fuses the scores of the three dimensions by setting weights, and finally outputs PLI in the range of 0–1 and various sub-indicators. This dual framework of density and coherence takes into account both information volume and sequence rationality, integrates multi-source features such as NER, geographical anomalies, and logic dictionaries, and enables efficient batch calculation without manual annotation, providing an objective and interpretable quantitative benchmark for comparing plots in large-scale tourism short videos.

2.2.2 Perceived Authenticity

This study constructs a measurement framework for perceived authenticity based on dual visual and semantic channels, systematically integrating video features such as frame color and contrast of short videos with the density of information-guiding words and imagery rhetorical

words in subtitle texts. By quantifying the degree of real-scene restoration at the visual level and the orientation between practicality and decoration at the semantic level, respectively, scores for the two dimensions of Genuine Authenticity and Alterreal Authenticity are obtained. On this basis, a normalized fusion method is adopted to map the above dual scores into a single comprehensive indicator, realizing the continuous measurement of perceived authenticity in tourism short videos. This indicator ranges from 0 to 1, with higher scores indicating that the video content is closer to real representation, and lower scores reflecting stronger artistic processing and Alterreal Authenticity characteristics. The above quantification scheme considers multimodal feature fusion, interpretability, and batch processing efficiency, providing a scientific and reliable measurement tool for empirical research on the Genuine Authenticity-Alterreal Authenticity continuum in tourism short videos

2.2.3 Content Value Orientation

Based on grounded analysis, this study divides the value dimensions of tourism short videos into two basic orientations: Hedonic Orientation, which focuses on emotional arousal and atmosphere creation, and Utilitarian Orientation, which emphasizes practical information and action guidance. To distinguish between the two in massive samples, the model takes subtitle semantics, domain keywords, text structure signals, and key visual features as multi-source inputs, forms a continuous orientation score ranging from 0 to 1 through hierarchical fusion, and outputs binary labels according to verified thresholds. This framework captures implicit emotions through deep semantic embedding while retaining interpretability using expert dictionaries and statistical features. It is robust to complex expressions such as short texts, slang, and negations. It provides a reliable quantitative benchmark for comparative studies between emotional mobilization and information services in tourism content.

2.3 Variable Selection

2.3.1 Agent Perspective

Agent Perspective refers to the subject position adopted in the linguistic expression of tourism short videos, including first-person perspective (e.g., "I arrived here"), second-person perspective (e.g., "You should come and see"), third-person perspective (e.g., "They are visiting"), and non-personal perspective (no apparent subject reference, only conveying information through images or text). Agent Perspective affects users' understanding of the "I-you-he" role relationships in the content and is an important linguistic mechanism for constructing cognitive engagement and psychological distance. The first person easily establishes a sense of identity engagement, the second person enhances interactive directivity, the third person tends to be objective in narration, and the non-personal perspective downplays subject intervention, emphasizing more on the event itself or the presentation of landscapes. Therefore, as one of the key dimensions of narrative language, Agent Perspective is incorporated into the model to explain its influence.

2.3.2 Cinematic Perspective

Cinematic Perspective refers to the contextual source and dialogue structure of the information narrated in the video, specifically including character self-narration (characters telling their own experiences), third-party narration (explanations by non-character narrators), multi-person dialogue (alternate expressions between characters), and non-personal presentation (no language input, only conveyed through images and background sounds). Different presentation methods mobilize the audience's cognitive processing paths and interactive perceptions. Character self-narration

strengthens intimacy, third-party narration enhances authority, multi-person dialogue constructs an authentic interactive atmosphere, and non-personal presentation focuses on scene atmosphere and intuitive immersion. The presentation method affects the comprehensibility of the narrative.

2.3.3 Viewport Perspective

Viewport Perspective refers to the aspect ratio structure of short videos on the terminal screen, mainly divided into three types: vertical screen (9:16), horizontal screen (16:9), and square screen (1:1). Viewport Perspective affects the way users focus visually during viewing and the adaptability to usage scenarios. Vertical screens are more suitable for mobile immersive experiences, horizontal screens are beneficial for wide-angle landscapes and spatial expansion, and square screens pursue picture balance and compositional aesthetics. The video presentation ratio determines the density of frame information and visual pressure, affecting viewing fluency and the willingness to stay. Therefore, this study incorporates it as a media form variable into the model.

2.3.4 Character Perspective

Character Perspective refers to the position from which the camera shoots, i.e., "from whose angle to look", including first perspective (from the subjective angle of the shooter), second-person perspective (the camera is directed at the audience, forming interaction), third-person perspective (observing others' behaviors from an outsider's angle), and non-personal perspective (no apparent character dominance in the frame). Character Perspective is related to the social distance and psychological projection mode between users and content. The first perspective enhances presence and immersion, the second perspective strengthens interactive experience, the third perspective emphasizes observability and narrativity, while the non-personal perspective highlights the atmosphere of the scene itself.

2.3.5 Rhythm

Rhythm refers to the temporal coordination and rhythmic matching between content, images, and background music in tourism short videos, reflecting the work's dynamic fluency and sensory rhythm. As a cross-modal narrative organization method, rhythm affects the audience's attention maintenance and cognitive resource allocation by regulating the speed and intensity of information output (Lang et al., 2007). Fast-paced videos are more likely to stimulate sensory stimulation and arousal, while a stable rhythm helps maintain emotional stability and sustained viewing. In tourism short videos, rhythm reflects the editor's and producer's style and affects the audience's sense of presence and immersion in the experience of scenic spots. Therefore, it plays a key role, so the article includes rhythm in the explanatory variables for analysis.

2.3.6 Structure

Structure refers to the organizational pattern of narrative arrangement in tourism short videos, mainly including three types: Unit Style (developed around a single scene or theme), Continuous Style (with sequential connection in time or space), and Discrete Style (composed of multiple relatively independent segments spliced together). Structure determines the audience's cognitive path to content coherence, and different structural organizations may affect the efficiency of information processing and the difficulty of plot acceptance. The continuous style helps establish a complete travel context and enhances the sense of plot progression; the unit style is more concentrated and suitable for expressing single-point experiences; and the discrete style provides diversity and rhythm.

2.3.7 Genre

The genre of tourism short videos refers to the overall structural form presented in their

narrative contexts and emotional evolution. Specifically, it can be divided into the following six types: Flash-Compilation Content, Emotive Expression Content, Gastronomic Exploration Content, Itinerary-Oriented Content, Cultural Interpretation Content, and Experiential Narrative Content. Flash-Compilation Content is characterized by quickly switching shots and highly rhythmic editing, emphasizing visual impact and rapid emotional guidance. Emotive Expression Content focuses on conveying profound emotional changes through language or images, usually manifested as emotional fluctuations and releases, primarily arousing the audience's emotional resonance. Gastronomic Exploration Content showcases local cuisine and catering culture, emphasizes the presentation of food and tasting experiences, and often attracts the audience's attention through detailed explanations and exquisite images. Itinerary-Oriented Content provides travel plans, attraction recommendations, and practical information, presenting a content structure with strong logic and high practicality. Cultural Interpretation Content emphasizes delivering the history, culture, and background knowledge of destinations, enhancing the audience's sense of knowledge acquisition through professional narrative methods. Experiential Narrative Content leads the audience to feel unique experiences in the travel process through the creator's personal experiences, focusing on presenting personal perspectives and subjective experiences. These six genres represent the most common narrative modes in tourism short videos, and each genre reflects the creators' different strategic choices in arranging content structures and promoting emotions. This study takes genre as an explanatory variable because, as a macro type of narrative structure, it significantly regulates the rhythm, guidance, and emotional participation of users' viewing behaviors. Different genres directly affect the audience's viewing paths by stimulating different psychological expectations and viewing experiences, determining whether the audience stays in the video and completes watching. Especially in the short video environment with high information density and limited time, genre, as an overall narrative framework, determines the attractiveness of Content and the way of emotional retention. Therefore, genre is regarded as an important variable in ND in this study.

2.3.8 Drama

Drama refers to the emotional tension and narrative ups and downs embodied in the plot design of tourism short videos, specifically including four types: Curiosity-Order Drama, Surprise-Order Drama, no dramatic structure, and the structure where curiosity and surprise coexist. Among them, Curiosity-Order Drama creates expectations by delaying information disclosure, such as setting unfinished or upcoming exciting scenes; Surprise-Order Drama breaks expectations through sudden reversals, such as unexpected endings or process reversals; no dramatic structure refers to the content being presented straightforwardly with stable emotional clues; while the structure where curiosity and surprise coexist integrates gradual foreshadowing and sudden turns, forming a more substantial emotional promotion effect. The above structures collectively constitute the dramatic level of the video in terms of sensory rhythm. Drama is included as an explanatory variable because different dramatic structures directly affect the degree of users' attention maintenance and emotional investment. Content with high Drama is often more likely to arouse the audience's expectations and viewing curiosity, prompting them to stay in the entire video until they obtain emotional release or information response. In the context of short tourism videos, Drama enhances the narrative appeal of the content. It guides the audience through plot climaxes and emotional focal points to form deeper viewing stickiness. Therefore, Drama is incorporated into this research framework as a key variable in ND.

2.3.9 Linguistic Style

Linguistic Style refers to the expression methods and communication paths presented in tourism short videos (Zhu et al., 2024), specifically including three types: Functional–Declarative, Affective–Narrative, and Imagistic–Semiotic styles. Among them, the Functional–Declarative Style focuses on practical information, emphasizes an objective and concise tone of statement, and is generally common in attraction introductions, transportation guidelines, and price explanations; the Affective–Narrative Style strengthens subjective experiences and audience resonance through individualized narration and emotional expressions; while the Imagistic–Semiotic Style does not rely on linguistic texts or voice expressions, but conveys atmosphere and values through non-linguistic symbols such as frame composition, rhythmic editing, and background music, with strong metaphorical and artistic qualities. As one of the core components of ND, Linguistic Style directly shapes the communication context and emotional tension of the content, and different styles will stimulate different cognitive processing modes and psychological expectations of the audience: practical language may promote goal-oriented viewing motivation, emotional language is more likely to form emotional investment and immersive experience. At the same time, non-verbal imagistic expression guides the audience to construct their meaning projection through sensory stimulation. In the communication context of short tourism videos, linguistic Style affects the audience's acceptance and understanding of the content. It has a potential guiding effect on retention time and behavior transformation during viewing. Therefore, it is included in the study as an important ND variable for examination.

2.3.10 Plot Logic

Plot Logic refers to the orderliness of event development and the rationality of narrative structure in travel short videos, which is mainly reflected in the internal consistency of Journey Progression, Scenic Transposition, and Causal Logic. Journey Progression focuses on whether events unfold in a realistic or comprehensible chronological order; Scenic Transposition emphasizes whether scene transitions have coherent geographical or spatial relationships, avoiding abrupt jumps; Causal Logic refers to whether there are clear causal connections and progressive logic between events within the video, ensuring the comprehensibility and persuasive power of the content. Plot Logic is employed as an explanatory variable to explore how the structural rationality of NC influences viewers' cognitive processing. A logically coherent plot structure helps viewers quickly construct situational models in a short time, reduces cognitive load, and enhances information absorption and willingness to watch; conversely, plot jumps or structural confusion may lead to cognitive disruption and loss of interest, and in travel short videos in particular, they are more likely to diminish the recognizability and sense of immersion in the destination's image. Therefore, this study incorporates Plot Logic as one of the key variables of NC.

2.3.11 Perceived Authenticity

Perceived Authenticity refers to the audience's subjective perception of Authenticity formed by the content presented in tourism short videos, mainly Genuine Authenticity and Alterreal Authenticity. Genuine Authenticity emphasizes a high degree of consistency between the content and the real world, such as whether the images are not exaggerated and whether the scenes conform to actual travel experiences. At the same time, Alterreal Authenticity refers to content that, although not entirely true, can stimulate a credible perception and emotional resonance in situational construction, for example, enhancing emotional tension through mild filters and editing without compromising overall credibility. Together, they form the basis of the audience's psychological

judgment of a credible yet engaging sense of reality, serving as an inseparable narrative mechanism in content perception. Perceived Authenticity is adopted as an explanatory variable because the viewing experience of tourism short videos relies on information acquisition and is influenced by the audience's subjective judgment of whether the content is credible and close to life. High perceived Authenticity helps establish a trusting relationship between the audience and the content, enhances viewing immersion and emotional investment, and increases the probability of SV. Whether it is the authentic presentation of daily travel or the idealized scenes constructed through alterreal strategies, the perception of Authenticity plays an important regulatory role in users' willingness to stay and continue watching, thus being included in the NC dimension for systematic investigation in this study.

2.3.12 Content Value Orientation

Content Value Orientation refers to the core value tendency embodied in the content presentation of short tourism videos, which mainly includes two dimensions: Hedonic Orientation and Utilitarian Orientation. Among them, Hedonic Orientation emphasizes the content's appeal to the audience in terms of sensory pleasure, emotional infection, and aesthetic experience, satisfying their psychological needs for entertainment and emotional regulation; Utilitarian Orientation focuses on the instrumental value of the content in knowledge transmission, travel suggestions, route planning, etc., serving the audience's practical decision-making and destination cognition. These two orientations constitute the basic types of the content value system of tourism short videos, corresponding to two viewing paths: emotional motivation and rational motivation. Content Value Orientation is an explanatory variable that explores how different value orientations affect users' SV. In the short video environment, the audience not only seeks information satisfaction but also desires emotional connection and immediate rewards, and whether the content can match their current psychological expectations will directly determine whether they continue to stay and watch. Especially in tourism short videos, works that balance emotional and practical values are often more attractive and can trigger stronger recognition and stickiness. Therefore, as an important psychological variable reflecting users' perceived benefits, Content Value Orientation is included in the NC.

Table 2 Measurement Explanations of Variables

Variable Type	Variable Name	Measurement
Narrative Discourse	Agent Perspective	The proportion of the most frequently used personal pronoun in the video
	Cinematic Perspective	The audio and visual content in the video, which involves detecting the number of speakers and the status of audio-visual synchronization in the video
	Viewport Perspective	Video Aspect Ratio
	Character Perspective	Detecting human faces, human bodies, and hand movements
	Rhythm	Beat, Editing Density, Motion Density, and Audio-Visual Synchronization
	Genre	Integrating video, audio, and text features
	Drama	Text sentiment analysis, visual and audio feature extraction

	Linguistic Style	Judging based on the proportion of functional, emotional, and imagistic key words and sentences in the subtitles
	Structure	The average similarity of video segments
Narrative Content	Plot Logic	Comprehensive evaluation of temporal, spatial, and logical dimensions based on visual and textual content
	Perceived Authenticity	The color features of video frames and the keyword density of subtitle content
	Content Value Orientation	Extract the visual and textual features of the video, combined with emotional and informational keywords

Appendix 3. Description of Experimental Materials

The experimental stimuli were developed through secondary production using Nanji Island (Wenzhou, China) as a unified destination setting. The destination was selected due to its moderate public recognition, which minimizes participants' prior knowledge and stereotypical expectations, ensuring that responses could be attributed primarily to the manipulations of plot logic (PLI) and rhythm (NRI) rather than destination familiarity. The original footage was sourced from the creator *Crane Jiayang*, available at https://www.bilibili.com/video/BV1qbbWzQELd. Based on this material, narration scripts, shot sequencing, and music rhythm were systematically adapted. The narration was recorded using AI voice synthesis that simulated the original speaker's tone to maintain consistency in timbre and emotional tone. All scripts were reviewed by researchers in tourism management to ensure conceptual validity and manipulation accuracy. Following a 2 (PLI: high vs. low) × 2 (NRI: fast vs. slow) factorial design, four customized short videos were produced:High-PLI × Fast-NRI, High-PLI × Slow-NRI, Low-PLI × Slow-NRI, and Low-PLI × Fast-NRI. Manipulation of Plot Logic (PLI): The *high-PLI* versions strengthened temporal sequencing by including explicit time markers (e.g., "then," "the next day," "in the morning"), clear causal linkages between key narrative nodes, and spatial or route cues to facilitate coherent story reconstruction. The *low-PLI* versions deliberately reduced such cues by omitting time markers and causal connectors, weakening spatial continuity, and using slightly abrupt transitions, while keeping all factual information constant across versions. Manipulation of Rhythm (NRI): The *fast-NRI* versions adopted a slightly accelerated narration speed (≈1.1×), increased editing density, shortened average shot duration, and used high-tempo background music (BPM > 100) to enhance audiovisual synchrony. The *slow-NRI* versions maintained normal narration speed, lengthened shot duration, and incorporated low-tempo music (BPM < 100) to evoke a more relaxed narrative flow. The BPM thresholds were established with reference to Ding and Lin (2012), and pragmatically adjusted for short-video contexts (above vs. below 100 BPM) to balance operability and perceptual discriminability. Before the subjective manipulation check, an objective calibration was conducted using the validated PLI and NRI quantification codes developed in Substudy 3, following the same computational parameters. The resulting indices for the four conditions were as follows:

- High-PLI × Fast-NRI: PLI = 0.645, NRI = 0.376
- High-PLI × Slow-NRI: PLI = 0.638, NRI = 0.322
- Low-PLI × Slow-NRI: PLI = 0.633, NRI = 0.339
- Low-PLI × Fast-NRI: PLI = 0.639, NRI = 0.409

A stratified comparison revealed consistent differentiation patterns: Within the fast-rhythm condition, the high-PLI video (0.645) exceeded the low-PLI video (0.639) by 0.006; within the slow-rhythm condition, the high-PLI video (0.638) exceeded the low-PLI video (0.633) by 0.005—both consistent with theoretical expectations. Along the rhythm dimension, under high-PLI conditions, fast rhythm (0.376) was higher than slow rhythm (0.322), with a difference of 0.054; under low-PLI conditions, fast rhythm (0.409) exceeded slow rhythm (0.339), with a difference of 0.070. These results indicate sufficient separation along both dimensions: rhythm manipulation exhibited clear and stable differentiation, while plot logic manipulation showed smaller but theoretically consistent separation. Considering these objective indices together with the significant results of the subsequent subjective manipulation check and the quantitative release

standards established during the pilot phase, all four stimuli were deemed valid and ready for experimental deployment. The corresponding video links are as follows:

- High-PLI × Fast-NRI: https://www.bilibili.com/video/BV1Koe1zrEWx
- High-PLI × Slow-NRI: https://www.bilibili.com/video/BV11oe1z6Eaq
- Low-PLI × Slow-NRI: https://www.bilibili.com/video/BV12de1z3EJB
- Low-PLI × Fast-NRI: https://www.bilibili.com/video/BV11oe1z6Esf

Appendix 4: Experimental Data

Table 3 Demographic Information

Variable	Main Experiment (n = 298)	
Gender		
Female	199	66.78%
Male	99	33.22%
Age		
Under 18	0	0.00%
18–25	94	31.54%
26–30	75	25.17%
31–40	97	32.55%
41–50	21	7.05%
51–60	9	3.02%
Above 60	2	0.67%
Occupation		
Student	47	15.77%
State-owned enterprise	155	52.01%
Public institution	48	16.11%
Government employee	21	7.05%
Private enterprise	8	2.68%
Foreign enterprise	15	5.03%
Other	4	1.34%
Education level		
Junior high school or below	2	0.67%
High school / Technical secondary school	8	2.68%
Associate degree	33	11.07%
Bachelor's degree	218	73.15%
Postgraduate and above	37	12.42%

Table 4 Measurement Items

Variable	Code	Item	Source
Narrative Transportation (NT)	NT1	I was completely immersed while watching this video.	Cao et al. (2021)
	NT2	I could vividly imagine the scenes depicted in the story.	
	NT3	While watching, I felt as if I were physically present in the destination shown in the video.	
Destination Cognition (DI)	DI1	This short video enhanced my understanding of the destination.	Beerli and Martin (2004)
	DI2	This short video helped me identify	

		the destination's core attractions and distinctive features.	
	DI3	I can distinguish key informational differences between this destination and others.	
	DI4	I can clearly summarize what to see, what to do, and how to get to this destination.	
Affective Attitude (AA)	AA1	Imagining a trip to this destination would be enjoyable.	Lam and Hsu (2004)
	AA2	Traveling to this destination would be a positive choice.	
	AA3	Visiting this destination would be interesting.	
	AA4	A trip to this destination would be pleasant.	
	AA5	Overall, visiting this destination would be a good idea.	
Travel Intention (TI)	TI1	I am willing to visit this destination in the future.	Wang et al. (2022)
	TI2	I would recommend this destination to others.	
	TI3	I plan to include this destination in my travel list.	

Table 5 Model Fit Indices

Index	CMIN/DF	GFI	AGFI	RMSEA	CFI	NFI	IFI
Ideal value	＜3	＞0.9	＞0.9	＜0.08	＞0.9	＞0.9	＞0.9
Acceptable threshold	＜5	＞0.8	＞0.8	＜0.10	＞0.8	＞0.8	＞0.8
Observed value	1.285	0.953	0.933	0.031	0.988	0.950	0.988

Table 6 Reliability and Validity Test

Variable	**Item Code**	**Reliability (α)**	**Factor Loading**	**CR**	**AVE**
NT	NT1	0.904	0.848	0.905	0.760
	NT2		0.849		
	NT3		0.869		
DI	DI1	0.859	0.808	0.860	0.606
	DI2		0.758		
	DI3		0.801		
	DI4		0.781		
AA	AA1	0.823	0.650	0.824	0.484

	AA2		0.746		
	AA3		0.735		
	AA4		0.780		
	AA5		0.694		
TI	TI1	0.718	0.843	0.739	0.494
	TI2		0.526		
	TI3		0.851		

Table 7 Discriminant Validity

Variable	1	2	3	4
1.NT	**0.872**			
2.DI	0.662***	**0.778**		
3.AA	0.401***	0.449***	**0.696**	
4.TI	0.252***	0.286***	0.617***	**0.703**

Note: The bolded diagonal values represent the square roots of the AVE for each construct.
p < 0.001.

References

Araujo, T., & Lock, I., & van de Velde, B. (2020). Automated Visual Content Analysis (AVCA) in Communication Research: A Protocol for Large Scale Image Classification with Pre-Trained Computer Vision Models. *Communication Methods and Measures*, 14(4), 239-265. Retrieved from https://doi.org/10.1080/19312458.2020.1810648. https://doi.org/10.1080/19312458.2020.1810648.

Beerli, A., & Martin, J. D. (2004). Factors influencing destination image. *Annals of Tourism Research*, 31(3), 657–681.

Cao, X., & Qu, Z., & Liu, Y., & Hu, J. (2021). How the Destination Short Video Affects the Customers' Attitude: The Role Of Narrative Transportation. *Journal of Retailing and Consumer Services*, 62. https://doi.org/10.1016/j.jretconser.2021.102672.

Ding, C. G., & Lin, C.-H. (2012). How does background music tempo work for online shopping? *Electronic Commerce Research And Applications*, 11(3), 299-307. Retrieved from https://www.sciencedirect.com/science/article/pii/S1567422311000627. https://doi.org/https://doi.org/10.1016/j.elerap.2011.10.002.

Gu, X., & Tse, C.-S. (2016). Narrative perspective shift at retrieval: The psychological-distance-mediated-effect on emotional intensity of positive and negative autobiographical memory. *Consciousness and cognition*, 45, 159-173. Retrieved from <Go to ISI>://WOS:000385054000014. https://doi.org/10.1016/j.concog.2016.09.001.

Han, M., & Yang, L., & Jin, X., & Feng, J., & Chang, X., & Wang, H. (2024, 16-22 June 2024). *Video Recognition in Portrait Mode.* Paper presented at the 2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR).

Lam, T., & Hsu, C. H. C. (2004). Theory of Planned Behavior: Potential Travelers from China. *Journal of Hospitality & Tourism Research*, 28(4), 463-482. https://doi.org/10.1177/1096348004267515.

Lang, A., & Byungho, P., & N., S.-J. A., & D., W. B., & and Wang, Z. (2007). Cognition and Emotion in TV Message Processing: How Valence, Arousing Content, Structural Complexity, and Information Density Affect the Availability of Cognitive Resources. *Media Psychology*, 10(3), 317-338. Retrieved from https://doi.org/10.1080/15213260701532880. https://doi.org/10.1080/15213260701532880.

Nan, X., & Futerfas, M., & Ma, Z. (2017). Role of Narrative Perspective and Modality in the Persuasiveness of Public Service Advertisements Promoting HPV Vaccination. *Health Communication*, 32(3), 320-328. Retrieved from <Go to ISI>://WOS:000392839800007. https://doi.org/10.1080/10410236.2016.1138379.

Wang, L., & Guo, Z., & Zhang, G.-y., & Xu, X. a. (2022). Effective destination user-generated advertising: Matching effect between goal framing and self-esteem. *Tourism Management*, 92, 104557. https://doi.org/10.1016/j.tourman.2022.104557.

Zhu, Z., & FANG, X., & Shan, M., & Chen, H. A. (2024). The effect of language style on consumers' perceived usefulness of online reviews: A regulatory focus-based study. *Nankai Business Review*, 27(03), 234-246. https://doi.org/12.1288.F.20230606.1625.004.