Lingue e Linguaggi
Lingue Linguaggi 74 (2026), 87-106
ISSN 2239-0367, e-ISSN 2239-0359
DOI 10.1285/i22390359v74p87-106
https://ese-journals.unisalento.it, © 2026 Università del Salento

HUMOUR AND MODAL INTERPLAY IN TIKTOK'S POV GENRE

AUDREY CLAIRE WILLOUGHBY
UNIVERSITÀ DEGLI STUDI DI MILANO LA STATALE

Abstract - TikTok is widely recognised for its humour-driven performances, yet the specific multimodal strategies through which humour is produced remain underexplored. This study examines humour in the platform's "POV" genre, analysing 16 English-language videos tagged with humour-related hashtags. Drawing on multimodal discourse analysis (Bateman *et al.* 2017; Wildfeuer 2014) and pragmatic humour theory (Attardo 2020; Tsakona 2020), it introduces *cross-modal incongruity* as a distinct form of modal interplay: a moment when one semiotic mode subverts the interpretive frame set by another, prompting the audience to reconcile the mismatch through inference. Six semiotic modes were examined to identify how expectations are established and violated. Cross-modal incongruity was present in every video analysed, most often when textual or verbal framing was overturned through embodied or visual cues. These mismatches acted as prompts for reinterpretation rather than signs of communicative breakdown, drawing on viewers' familiarity with genre conventions, platform performance styles, and shared cultural scripts. The findings suggest that such incongruities are systematically embedded in the multimodal design of POV videos, making the audience's interpretive work integral to the humour and offering a clearer account of how comic effects are orchestrated across audiovisual resources in short-form digital media.

Keywords: multimodal discourse analysis; TikTok; pragmatics; humour; incongruity.

1. Introduction

Short-form video platforms such as TikTok are reshaping how humour is produced and interpreted in everyday digital communication. In highly compressed audiovisual formats, creators draw on semiotic resources such as text overlays, voice, soundtracks, facial expression, gesture, and editing techniques to simulate familiar scenarios and frame social roles. Humour often emerges when these modes clash or shift unexpectedly, producing moments of dissonance that disrupt interpretive coherence and invite incongruous readings (Darvin 2022).

From a pragmatic perspective, humour is not simply a property of a text but a communicative act whose interpretation depends on the interplay between what is said, how it is said, and the shared knowledge and

expectations of participants (Attardo 2020; Tsakona 2020). In this view, incongruity, or violation of expectation, becomes a source of humour through contextualised inferencing, social positioning, and genre competence, which determine whether a multimodal cue is interpreted as signalling the humorous nature of the text.

In TikTok's "point of view" (POV) genre, creators present short fictional scenarios from a first-person or implied perspective, often dramatising everyday dynamics such as parenting, school life, or friendship. These performances rely on rapid sequencing of modes, such as textual overlays, gestures, borrowed media, and audio, where humour often arises from incompatibility in modal alignment that prompt reinterpretation (Bernad-Mechó and Girón-García 2023; Masi 2023). These interferences occur in a temporal sequence, as discussed more below.

While humour in TikTok's POV genre videos often relies on modal shifts that disrupt audience expectations, little research has examined how these violations are constructed and are temporally distributed in short-form audiovisual media. This study addresses that gap by combining multimodal discourse analysis with pragmatic humour theory to investigate cross-modal incongruity as a humour-making strategy, analysing its development over time, across modes, and within narrative positions in relation to contextual expectations.

2. Literature review

2.1. Platform affordances and the POV genre

TikTok's POV genre is shaped by a combination of platform affordances, established performance practices, and platform-specific conventions. In many ways, POV videos build on the logic of static-image memes: while memes condense humour into a single frame, POV videos extend these memetic structures into dynamic, temporally sequenced performances shaped by TikTok's affordances (Shifman 2013; Zulli and Zulli 2020). Traditionally, POV videos were visually framed through the eyes of the character viewers were invited to embody. This literal, first-person usage was present on TikTok from its earliest examples. However, by 2019, the term had broadened to describe videos that invite viewers to engage with a character's perspective or experience without adopting their literal, visual vantage point (Attardo 2025). Both uses have coexisted since that time. In this study, this expanded sense is referred to as the "theatrical" POV, in which scenarios unfold visibly on screen, encouraging relational or emotional identification rather than immersive embodiment.

Platform affordances such as green screen effects, voice dubbing, text

overlays, and audio synchronisation layer meaning: textual overlays can establish roles or narrative frames, while facial expression, camera movement, and sound design set tone and pacing (Divon and Krutrök 2024). As Bucher and Helmond (2018) argue, affordances should not be seen merely as technical features but as action potentials that emerge through interaction between users and the socio-technical environment. On TikTok, these tools enable compact, stylised narratives that draw on cultural scripts to perform irony, exaggeration, or sincerity (Kułaga 2024; Zeng and Abidin 2021; Zulli and Zulli 2020).

The POV format thrives on ambiguity: it gestures toward shared experience while also inviting playful interference. As Trillò (2024) describes, POV videos participate in TikTok's broader "platform vernacular", where repetition, remix, and stylised stance-taking establish community belonging (p. 3). The humour in TikTok's POV genre frequently emerges from viewers' recognition of a familiar interactional script, such as a teacher scolding a class or a couple arguing over something seemingly mundane, and the subtle ways this script is manipulated across modes. Textual overlays frequently establish the expected frame of reference, signalling a character type or scenario, while the subsequent performance introduces dissonance or exaggeration that challenges the viewer's initial assumptions. This approach reflects what Trillò (2024) identifies as a mode of performance in which creators draw on shared cultural scripts to signal affiliation with their audience, while leaving space for irony and playful reinterpretation. Similarly, Divon and Ebbrecht-Hartmann (2022) note that POV videos often organise their narratives and emotional cues through a coordinated use of visual framing, facial expression, and direct camera address, creating a multimodal structure that shapes how viewers interpret and respond to the performance. As Aiello and Parry (2020) explain, these types of shifts are made intelligible through meta-communicative cues that guide interpretation and indicate stance – a process that closely aligns with Gumperz's (1982) notion of contextualization cues, which function similarly to signal speaker stance and frame interpretation as the discourse unfolds.

This use of shared cultural scripts is reinforced by a reliance on memetic structure and repetition, which also contributes to the humour. Shifman (2013), Zulli and Zulli (2020), Vásquez (2019), and Attardo (2023) argue that the viral spread of digital genres is often rooted in repetition-with-variation: creators replicate a known format while introducing subtle shifts in delivery, stance, or outcome. The POV genre benefits from this logic. Users internalise patterns and remix them, signalling affiliation while also inserting playful contradictions. Ironic editing, exaggerated delivery, and visual discontinuities all foster re-interpretation and shared recognition, making humour a product of both performance and platform literacy.

2.2. Pragmatic incongruity

Incongruity, typically defined as a clash between what is expected and what is presented, has long been recognised as central to humour (Attardo 1994, 2000; Raskin 1985). From a pragmatic perspective, incongruity functions as a communicative breach: a recognisable mismatch between surface form and inferred meaning (Attardo 2020; Gerrig 1984). In digital contexts, such breaches frequently occur across multiple modes reflecting broader shifts in humour toward fast, interactive, and multimodal formats (Attardo 2023).

This study adopts a pragmatic view of incongruity, where humour arises not only from contrast or contradiction, but also from a viewer's interpretation of intentional misalignment between cues. As Tsakona (2020) points out, humour is always contextually situated as it reflects shared knowledge, genre norms, and social positioning. In the POV genre, creators frequently perform exaggerated versions of familiar roles, using irony and over-performance to stage communicative breaches that may be interpreted as humorous precisely because they deviate from expectations in subtle, stylised ways.

Viewers are expected to infer humorous intent not necessarily from overt jokes or punchlines, but from friction between modes. The idea that humour may stem from unresolved or only partially resolved incongruity that leaves interpretive space open was noted by Attardo and Chabanne (1992) in their analysis of jokes as a text type, and has since been extended to static online formats by Dynel (2016) and Abdel-Raheem (2018), where humour frequently emerges from visual–verbal tension, particularly when image and text combine to set up and subvert expectations. In audiovisual digital communication, such tensions can arise when, for example, a textual overlay suggests a serious scenario that is undercut by an absurd soundtrack or a deadpan reaction. The incongruity lies in the modal misalignment, or in what Holsanova (2014) describes as attentional and interpretive shifts across various modes. These shifts often occur without explicit signalling, relying instead on the viewer's genre literacy and familiarity with platform-specific cues. As Yus (2018) observes, humour in these contexts hinges on the audience's inferential effort to detect relevance among potentially contradictory signs. When viewers recognise that a gaze, gesture, or tonal shift contradicts the expected function of a caption or verbal frame, they resolve this discrepancy through contextual reasoning. The process is dynamic and rarely results in narrative confusion; rather, it invites playful reinterpretation and aligns with the participatory nature of TikTok's platform vernacular. Recognising the breach requires cognitive attention and social competence, lending a sensitivity to genre, irony, and communicative norms.

Humour in POV videos is often built around semiotic violations that subtly interfere with coherence across modes. In this genre, such violations

operate much like the modal misalignments described above, but they are staged within highly compressed, stylised narratives. These moments invite viewers into an active interpretive role, requiring them to identify and reframe unexpected modal shifts as intentional.

2.3. Multimodal approaches to humour

Multimodal discourse analysis (MDA) has received relatively little attention in the study of humour in audiovisual genres, and work on informal, user-generated formats is even more scarce. In particular, few studies have examined the specific role of different modes in creating humorous effects in these contexts. Only recently has MDA been applied to short-form, socially embedded video platforms such as TikTok, where humour circulates in highly participatory and multimodal ways.

Most studies concerned with the modal interplay of online humour have centred on static image–text memes, where juxtaposed modes create contrasts through mismatched tone, framing, or logic (Abdel-Raheem 2018; Dynel 2016; Shifman 2014; Vásquez and Aslan 2021; Yus 2018). While these analyses reveal much about how humour can arise from visual–verbal tension, their focus on fixed, spatial arrangements means the sequence of interpretation is constrained by a single, unchanging visual-textual composition. By contrast, in audiovisual genres, cues from multiple modes unfold, build, and transform as the text progresses. Their sequential organisation actively reshapes expectations over time before they are playfully violated. These genres require analytical approaches that can account for how meaning develops through successive modal shifts, capturing both the temporal organisation of cues and their interaction in producing humour.

Some research has addressed the role of modes in humorous dynamic media: Balirano and Corduas (2008) analyse scripted diasporic television, showing how verbal dialogue, visual framing, and audio cues combine to produce humour rooted in shared cultural knowledge. Bernad-Mechó and Girón-García (2023) examine educational YouTube videos, tracing how humour emerges through the interplay of gesture, the pacing shaped by editing rhythms, and shifts in narrative perspective. While their analysis addresses timing, it centres on the overall structure of a performance rather than on the precise sequencing by which expectations are formed and overturned. Masi (2023) studies how speakers in TED talks use gaze, gesture, prosody, and visual design to create humorous moments through incongruity. Although the study offers valuable insight into multimodal coordination, it gives less attention to how these cues unfold over time to gradually establish and then subvert expectations. These works underscore the importance of orchestrating multiple modes in time yet leave room for closer examination

of the temporal processes through which expectation and violation are constructed.

3. Methodology

3.1. Theoretical approach

This study draws on the works of Bateman, Wildfeuer, and Hiippala (2017) and Wildfeuer (2014), with particular attention to their concepts of modal interplay, interpretive framing, and multimodal coherence. Both frameworks offer structured models for understanding how meaning is constructed across multiple semiotic modes in communicative artefacts.

Bateman *et al.* (2017) describe modal interplay as the ways in which different semiotic modes relate to one another in producing a coherent multimodal discourse. Three primary relations are distinguished in their work: alignment, where modes support the same rhetorical or informational goals; elaboration, where one mode expands, specifies, or adds detail to content presented in another; and contrast, where the content presented in each mode is recognisably different yet integrated into a unified whole. In all cases, coherence is maintained through the integration of modal contributions. Wildfeuer (2014) complements this perspective by emphasising interpretive framing and the viewer's active engagement, showing how filmic discourse guides inference and coherence through shifts in devices such as camera angle, character perspective, or sound design. This focus on the audience's inferential work grounds the present study's extension of both frameworks. Recent work by Zabalbeascoa and Attardo (2023) emphasised that humour should not be understood as confined to linguistic expression, but rather as emerging through semiotic resources more broadly, a shift that underscores the importance of accounting for meaning across modalities.

Building on these perspectives, the current analysis distinguishes modal contrast from *cross-modal incongruity* as possible modal relations that can serve as triggers for humorous interpretation. Whereas modal contrast (as conceptualised by Bateman *et al.* 2017) involves differences between modes that remain structurally coherent and fully integrated into the discourse, offering alternative perspectives or nuances without creating interpretive interference, cross-modal incongruity designates moments in which one or more modes undermine, contradict, or otherwise subvert the interpretive frame established by another. Such instances produce a temporary breakdown in cross-modal coherence, requiring the viewer to resolve the misalignment through pragmatic inference or to interpret it as nonsensical. In this study, only such marked violations of expectations between modes are identified as

cross-modal incongruities and examined as key points in the construction of humour.

This analytical approach is guided by two central research questions:

1. Is cross-modal incongruity central to the humour of TikTok POV videos, and what semiotic strategies are used to achieve it?
2. How are cross-modal incongruities distributed across modes, segments, and narrative positions in humorous POV TikTok videos?

3.2. Data collection

The dataset consists of 16 English-language TikTok videos retrieved via TikTok's API belonging to the "POV" genre, tagged with humour-related hashtags (e.g., #comedy, #funny). From an initial pool of 20 videos, four were excluded: three due to privacy changes and one for having removed the audio.

Metadata such as upload date, like count, hashtags, and captions were recorded but are not analysed in this study. All videos were uploaded between August 2023 and early 2025. Creator demographics were not consistently available and are therefore not addressed.

3.3. Segmentation method

To analyse how expectation and violation unfold to create cross-modal incongruities in TikTok's short-form videos, this study adopts a segmentation method based on interpretive framing (Wildfeuer 2014). Rather than relying on fixed units, such as scenes or time intervals, segmentation follows shifts in the viewer's interpretive orientation prompted by modal cues including gaze, gesture, sound, or camera perspective. Principles from event segmentation theory (Zacks and Swallow 2007) were also integrated, which explains how viewers identify boundaries in unfolding activity based on perceptual salience, such as changes in tone, motion, or visual layout.

A segment is defined here as a bounded stretch in which a stable interpretive frame holds. A new segment is triggered when a shift in one or more modes reorients the viewer's understanding of the communicative act. This framing-based approach is well-suited to TikTok's compressed, stylised format, where coherence depends on rapid multimodal cues and changes in tone or communicative intent.

Each segment is manually annotated in Microsoft Excel, with a new segment identified whenever a semiotic shift is judged to reorganise the viewer's perspective or communicative expectations. This segmentation method enables fine-grained tracking of how expectation is built and then violated across modes, forming the basis for identifying cross-modal

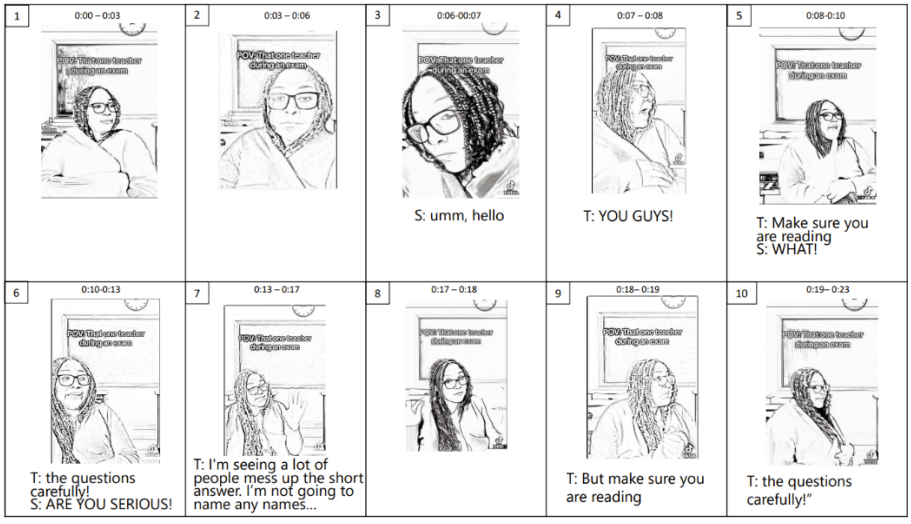
violations and capturing how humour is performed through the temporal interaction of modes.

3.4. Identifying cross-modal incongruity

Each segment was analysed in terms of six semiotic modes: written text, verbal spoken language, audio (such as music or sound effects), visual design (including background, editing choices, setting, and props), gesture, and facial expression (including gaze). For each mode, meaning was paraphrased, and segments were flagged for incongruity when they showed an interference in interpretive alignment either within a single mode or between modes. In these cases, the analysis identified which modes established the initial expectation (Meaning 1) and which violated said expectation (Meaning 2). Cross-modal incongruity was annotated when the mode or mode combination establishing the initial expectation differed from the mode or mode combination producing the violation, for example, when a joyful textual overlay was followed by a flat, ironic delivery, or when a facial expression contradicted the tone of spoken language. Incongruity could also occur within a single mode, such as a textual overlay that itself contained a tonal or logical mismatch. The aim was not to classify types of humour, but to identify recurring patterns in how audiovisual resources violate expectation.

3.5. Examples: Segmentation and cross-modal incongruity

3.1.5. Example one: V_001



Note. S: – POV student speech, T: – teacher speech

Figure 1
Video V_001.

Figure 1 demonstrates how a video from the data set was segmented. Video V_001 provides a representative case of how segmentation and violation of expectations are used to identify cross-modal incongruity. The video features a 23-second performance titled "POV: That one teacher during an exam", in which a single speaker enacts a teacher figure with exaggerated gestures and classroom mannerisms. Across ten segments, the video gradually builds an interpretive frame of a strict, mildly passive-aggressive authority figure through the interplay of textual overlay, gaze, gesture, and speech.

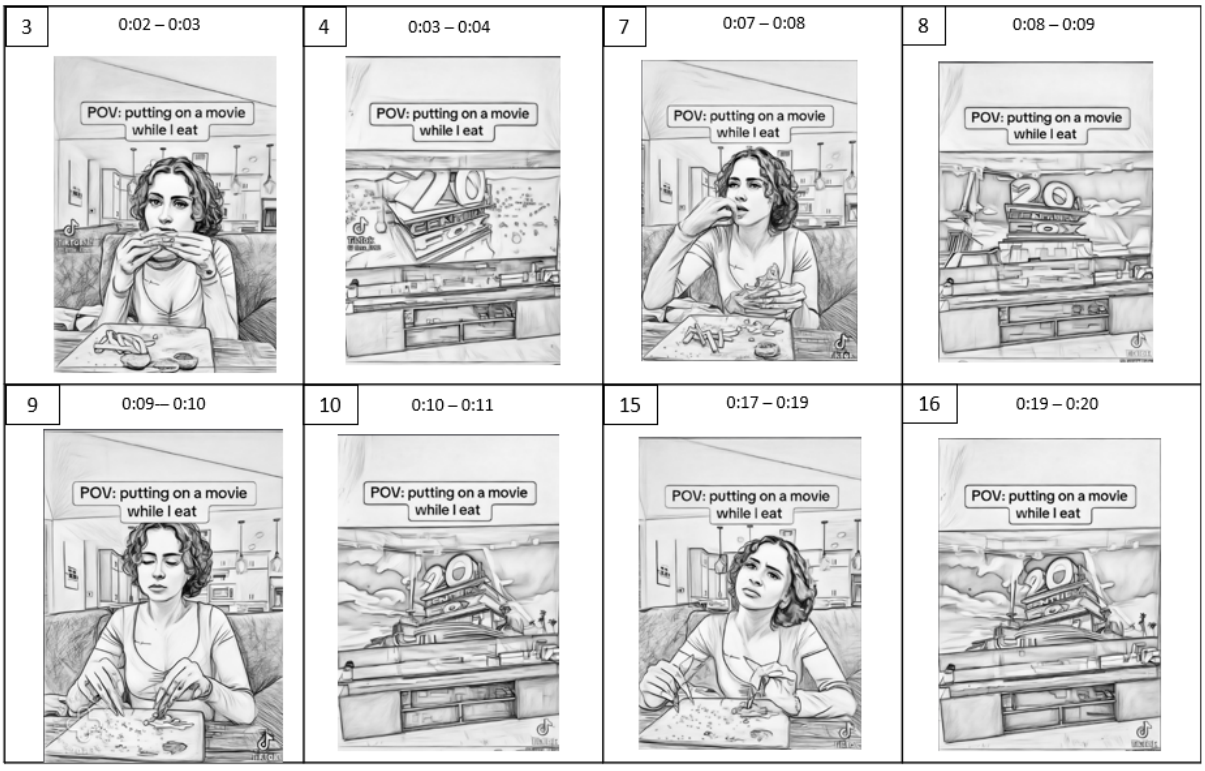
The first three segments are non-verbal: The video opens to the teacher scanning the classroom in silence with pursed lips as she crosses her arms, folding her sweater across her chest. At second 0.03 she tilts her head and makes eye contact with the camera/POV student, indicating a new segment. Around second 0:06 her attention turns to the POV student's work, as she also raises her eyebrows and upturns her mouth, indicating judgment and disapproval at what she sees, signalling the beginning of the third segment. The meaning-making across modes within these segments works in close alignment with the initial POV textual overlay by elaborating on one another to reinforce a consistent scenario, role, and tone.

Segments 4 through 7 each mark subtle reorientations. Segment 4 introduces the student's discomfort when he says, "Um, hello…", drawing attention to the teacher's continued silence. Segment 5 begins with the teacher's first spoken performance in the video, as she pulls away from the POV student and addresses the entire class, instructing everyone to double-check their work after viewing the POV student's exam. Segments 6 and 7 elaborate on this frame as she continues in a lecturing tone, with her index finger raised, to mildly scold about the performance of the test. These changes in speech act, gaze direction, and facial expression justify new segment boundaries, since perceptual shifts of this kind cue the viewer to register a boundary in unfolding activity (Zacks and Swallow 2007) and prompt an update in how the ongoing event is mentally represented. At the same time, they function as interpretive framing devices in Wildfeuer's (2014) sense, guiding the viewer's understanding of communicative intent.

A single, clear incongruity occurs across segments 7 and 8. In Segment 7, the teacher delivers the line: "I'm not gonna name names" with her hands raised, palms outward, as clearly seen in the corresponding visual frame in Figure 1. These verbal and gestural cues set an expectation of discretion, aligning with the role of a strict but fair authority figure. However, in Segment 8, she makes deliberate and prolonged eye contact with the camera (positioned as the POV student) during an awkward pause, hands still raised, thereby undermining the very expectation she just constructed by implying who the student-in-question actually is. This breach is not accidental or narratively repaired, but rather staged as an ironic, stylised contradiction between modes.

This is the only moment in the video that qualifies as a cross-modal incongruity. The humour arises from the viewer recognising the mismatch between verbal intent and gaze delivery. Rather than being explicitly resolved narratively, this incongruity remains open-ended, inviting pragmatic reinterpretation. In Segments 9 and 10, the teacher resumes her authoritative tone, returning to addressing the entire class without acknowledging the earlier interference. Although unresolved within the video's explicit narrative, viewers readily interpret this violation of expectation as humorous, recognising it both as a deliberate, exaggerated performance typical of the POV genre and as a familiar behavioural pattern. The humour, in other words, is not only triggered by the incongruity itself but also by the recognition of a familiar situation now being made visible and exaggerated for comic effect. This example thus demonstrates how segmentation based on modal shifts and interpretive reorientation can capture humour that emerges from the interplay of carefully timed intermodal contradictions and culturally shared recognition.

3.2.5. Example two: V_007



Note. Not all 16 segments of V_007 are present in Figure 2 due to space constraints; only 8 were selected to demonstrate the use of specific modes to convey the passage of time, as discussed below. The segment number out of the 16 total is listed in the upper left corner of each visual frame.

Figure 2
Video V_007.

Video V_007 is a 20-second montage of 16 segments with no spoken dialogue, alternating between shots of a young woman eating at a coffee table in her living room and shots of the modern 20th Century Fox film intro playing on her television. The actual length of the Fox intro is approximately 20 seconds, the same length as this TikTok. Yet through the montage structure, the viewer perceives it as lasting far longer: the woman's food visibly diminishes at a normal eating pace, creating the impression that more time has passed than the actual 20-second duration of the intro. The music of the intro plays continuously and at normal speed across cuts, meaning it is heard even when the camera cuts to the young woman eating, and the visual movements of the intro continue to unfold whenever the camera returns to the TV.

The video opens with the textual overlay "POV: putting on a movie while I eat", framing the scenario and establishing *meaning 1*: the expected sequence of eating a meal while a film plays. In segment 1, she bites into a burger, fries and chicken nuggets on the plate beside her. Segment 2 cuts to the Fox intro, its animation and music progressing normally as mentioned above. Segment 3 returns to her, taking a large bite; the burger is visibly smaller than it was in segment 1. Segment 4 cuts back to the Fox intro, still unfolding. This alternation continues: segment 5 shows her eating a chicken nugget; segment 6 again cuts back to the intro, progressing; segment 7 shows the burger now halved, the nuggets gone, and her focus shifting to the fries. Segment 8 again shows the intro's movements and music continuing. As this sequence develops, the embodied mode of eating and the visual cue of disappearing food begin to suggest that more time is passing between cuts than the real-time unfolding of the intro would allow, elongating its perceived duration.

By segment 9, most fries, and therefore almost the entire meal, are gone, and she glances down at her plate, followed by segment 10, in which the intro is still in motion. This is the point where *meaning 2* emerges: the expectation of enjoying a meal during a movie is violated because the meal ends before the film even begins. The incongruity could, in theory, also be read as her eating unusually fast. Still, because her on-screen eating is shown at a normal pace, inference favours the first interpretation - that the intro in the narrative reality of the TikTok is absurdly long.

Segments 11 and 12 reinforce the violation: the final bites of fries are eaten, the plate is empty, yet the intro continues unfolding on-screen. In segment 13, she wipes crumbs from her hands, an embodied cue of completion, followed by segment 14, yet another view of the still-progressing intro. Segment 15 delivers the reaction: she rests her arms on the table, tilts her head toward the TV, furrows her brow, and blinks in exaggerated disbelief. Segment 16 closes with one final cut to the intro as it continues still.

Embodied cues make the violation visible: the diminishing food portions, the hand movements signalling eating and then completion, and her closing reaction all contrast with the uninterrupted progression of the intro. The music's steady pace and the fluid motion of the intro confirm that no time compression or acceleration is happening within the intro itself. The montage thus overlays two timelines whose durations become incompatible in the viewer's mind: the completion of a meal versus the start of a film.

The humour here depends on the alignment and later divergence of modes. Textual framing, embodied action, and visual details of the disappearing meal set and sustain the initial expectation, while the visuals and audio of the 20th Century Fox intro establish the competing temporal frame. The violation lies in their sustained mismatch, a relatable recognition of how long movie intros can feel, reframed here through an exaggerated scenario.

As Attardo (2020) argues, the processing of incongruity depends on inferential mechanisms tied to what the viewer already knows about a character, context, and genre. In both examples, viewers recognise the contradictions not as communicative failures, but as deliberate cues for humour, interpreting them through familiarity with the POV format's exaggerated roles and performance conventions. Tsakona (2020) similarly emphasises that humour is shaped by genre conventions and social positioning, both of which inform how audiences read the teacher's performative breach in V_001 and the temporal absurdity of V_007. In each case, the incongruity remains unresolved within the narrative: in V_001, the teacher's direct gaze at the "unnamed" student is left hanging, while in V_007, the never-ending movie intro continues without closure. These deliberate non-resolutions work as signals to the audience that the contradiction itself is the point of amusement, encouraging them to actively interpret and engage with the performance.

4. Results and discussion

4.1. Cross-modal incongruity in the dataset

The analysis of 16 TikTok POV videos reveals that multimodal incongruity is a fundamental mechanism in how humour is constructed within this dataset. In each video, at least one instance was observed in which two or more semiotic modes collaboratively established an expectation, which was then violated another mode, creating a moment of interpretive interference. These instances were coded as occurrences of cross-modal incongruities and form the core of the analysis presented here.

4.2. Frequency and distribution of incongruity

A total of 28 instances of incongruity were identified across the 16-video dataset (see Table 1), 26 of which cross-modal incongruities, and 2 monomodal incongruities.

Video	Video length	# of segments	# of incongruity	Segment in which incongruity/ies occur(s)
V 001	00:23	9	1	7,8
V 002	00:15	6	3	1, 2, 6
V 003	00:10	2	2	1*, 1
V 004	00:07	4	1	1
V 005	00:10	2	1	2
V 006	00:14	6	1	3
V 007	00:20	16	2	9,10
V 008	00:19	9	3	4, 8, 9
V 010	00:07	4	1	3
V 011	00:11	7	2	2, 6
V 012	00:06	5	2	1*, 1
V 013	00:06	5	2	2, 4
V 014	00:07	2	1	1
V 016	00:14	4	1	1
V 017	00:06	2	1	1
0 019	00:14	10	4	5, 6, 8, 10

Note. Only 2 of the 28 instances of incongruity were monomodal, indicated with an astrik (*) in Table 1.

Table 1
Incongruity occurrences and segmentation across 16 videos in dataset.

All 16 videos in the dataset contained at least one instance of incongruity, distributed as follows:

- 8 videos featured one instance of cross-modal incongruity
- 3 videos featured two instances of cross-modal incongruity
- 2 videos featured three instances of cross-modal incongruity
- 1 video featured four instances of cross-modal incongruity
- 2 videos featured two instances of incongruity: one cross-modal and one monomodal textual incongruity

This distribution suggests that while many creators build toward a single key moment of modal interference, others layer multiple incongruities throughout the video, constructing more complex humorous arcs. Attardo's (2001) distinction between jablines and punchlines helps account for this variation: some incongruities operate like jablines, dispersed across segments as secondary humorous turns, while others resemble punchlines, concentrated in the final position as climactic resolutions. In this sense, the segmental positioning of incongruities becomes central to understanding how humorous effects are organised within POV clips.

The consistent presence of at least one instance of multimodal incongruity in every video further indicates that intermodal expectation

violation is a defining feature of humorous performance in the POV genre. Monomodal incongruities, while rarer, still contribute to the overall logic of humour in this genre by destabilising the established frame or subverting expectations within the same communicative channel.

4.3. Segmental positioning of incongruity

Occurrences of multimodal incongruity appeared at different narrative points but most frequently in the middle or later segments, following the establishment of a stable interpretive frame. For example:

- In V_001, cross-modal incongruity emerges across segments 7 and 8, where bodily and tonal shifts undermine the verbal premise
- V_008 presents three instances of cross-modal incongruity across segments 4, 8, and 9, suggesting a pacing strategy of escalating interference
- In V_011, a cross-modal incongruity occurs already in segment 2 out of 7, demonstrating that humour can also arise early in the video if enough contextual framing is established in the opening moments

In contrast, seven videos feature a cross-modal incongruity immediately in segment 1. In six of these, creators recontextualise intertextual media clips (remix), such as scenes from television shows, movies, or other viral videos, as the audiovisual base for their POV scenarios. These clips are paired with new textual overlays that assign an interpretive frame often at odds with the original content. The result is an immediate incongruity between the expectation created by the textual overlay and the meaning created by the audiovisual material.

For instance, in V_004, the textual overlay reads “POV: me trying to study”, setting up an expectation of studiousness and focused effort. In contrast, the accompanying video clip from *Ice Age 2* shows Sid the Sloth flailing on the ground and making guttural, nonsensical noises. The textual mode and the audiovisual clip belong to entirely different semantic domains: the text evokes discipline and concentration, while the moving image and sound convey chaos, absurdity, and lack of control. This stark mismatch produces an initial interference in interpretive alignment. Viewers resolve the interference by reframing the chaotic clip as a metaphor for mental disorganisation or the difficulty of maintaining focus – a familiar scenario to anyone who has ever tried to study – so that the incongruity ultimately functions as a source of humour.

These immediate contrasts rely less on narrative buildup and more on the viewer’s ability to resolve cross-modal incongruity in the moment, interpreting the misalignment through inference and cultural reference. In such cases, the humorous effect emerges not from progression, but from the

interplay between media recognition and textual framing.

4.4. Modal roles in expectation and violation

Each instance of incongruity was analysed by identifying which mode(s) contributed to the initial expectation (Meaning 1) and which violated it (Meaning 2). A breakdown of these interactions per video is shown in Table 2.

Video	Meaning 1 (Expectation)	Meaning 2 (Violation)
V_001	verbal	Expression
V_002	text	Visual
V_002	text	Verbal
V_002	text	verbal, visual
V_003	text	Text
V_003	text	Audio
V_004	text	audio, visual
V_005	gesture	audio, gesture
V_006	verbal	Expression
V_007	gesture, text, visual	audio, visual
V_007	gesture, text, visual	audio, visual
V_008	gesture	Expression
V_008	gesture	Expression
V_008	gesture, audio	expression, gesture
V_010	verbal	expression, gesture
V_011	verbal	Expression
V_011	verbal	Expression
V_012	text	Text
V_012	text	expression, visual
V_013	gesture, text	Audio
V_013	text	Gesture
V_014	text	Visual
V_016	text	verbal, visual
V_017	text	verbal, visual
V_019	visual	expression, gesture,
V_019	visual	expression, gesture
V_019	visual	audio, expression, gesture
V_019	audio, text, visual	audio, expression, gesture

Table 2
Expectation and Violation Modes per Video.

The distribution of modal roles across the 28 instances of incongruity reveals consistent patterns in how expectation and violation are constructed, outlined more clearly in Tables 3 and 4 below.

Mode	Occurrences
text	16
gesture	7
visual	6
verbal	5
audio	2
expression	0

Note. Occurrences across the entire data set.

Table 3
Meaning 1 – Expectations.

Mode	Occurrences
expression	13
visual	9
gesture	8
audio	8
verbal	4
text	2

Note. Occurrences across entire data set.

Table 4
Meaning 2 – Violations.

Textual overlays led as the primary framing mode to set up expectations that are subsequently violated (16 instances) as seen in Table 3. Gesture (7) and visual (6) modes were the second most common expectation-setting modes, often adding specificity to the scenario or emphasising the tone implied by the textual overlay. Verbal speech followed with 5 occurrences, typically reinforcing or elaborating the role or premise established in the overlay. Audio was comparatively rare in establishing expectations (2 occurrences), but when present, it often invoked recognisable media references or environmental cues that anchored the setting. These figures suggest that linguistic cues, particularly the POV textual overlay, operate alongside embodied and visual signals to construct the viewer's interpretive frame, signalling character identity, social context, or emotional tone, and preparing the viewer to read the scene in a particular way.

In contrast, Table 4 demonstrates that violations were most commonly enacted through expressive and embodied modes. Facial expression contributed to Meaning 2 in 13 instances, visual cues in 9 instances, and gesture and audio in 8 instances each. Verbal interferences occurred in 4 cases, while text appeared only twice, as part of a violation, and in both instances, it was the only monomodal incongruity in the dataset. The cross-modal incongruities in this dataset frequently took the form of embodied or visual interference to an established linguistic premise, creating a recognisable gap between what is verbally or textually implied and what is

shown or performed.

This recurring structure, in which linguistic framing is destabilised by visual or embodied modes, sets the stage for a broader theoretical reflection. It appears to support, at least in the context of POV humour, Lotman's (1975) proposal that language functions as a "primary modelling system" through which other semiotic systems derive coherence. In these examples, linguistic elements often establish the initial meaning configuration, while subsequent modalities introduce shifts that reframe or violate that configuration. By comparison, other TikTok genres, such as dance trends, visual memes, or reaction videos, may open with non-linguistic modes like movement, sound, or facial expression to set the premise. This variation in which mode initiates meaning shows that modal salience is shaped by genre, supporting multimodal discourse theories (Bateman *et al*. 2017; Jewitt 2014; Kress and van Leeuwen 2001) that view modes as operating through shared semiotic principles rather than presupposing inherent communicative dominance of one mode over others.

In the case of POV humour, linguistic and visual scaffolding is often deliberately set up so that it can be undermined. This aligns with Tsakona's (2020) view of humour as a pragmatic and context-dependent phenomenon shaped by shared genre expectations. Creators in this dataset draw on familiar social archetypes, such as teachers, parents, or work-related situations, which are rendered legible through text, gesture, or visual detail, and then exaggerated or subverted through facial expression, tone, or visual reversal. The segmentation method used in this study was instrumental not only in isolating these moments of cross-incongruity but also in capturing how they unfold over time and across different modes. This offers a practical model for identifying humour that is performed and modally layered, rather than explicitly stated.

5. Conclusion

This study shows that cross-modal incongruity is a deliberate and consistent humour-making strategy in TikTok's POV genre. In every video analysed, at least one expectation was established in one mode, then later violated by another, most often when textual or verbal framing was overturned through facial expressions, gestures, or visual details. These moments invited viewers to reinterpret the scene in light of the new cue, recognising the misalignment as intentional and drawing on genre knowledge, platform norms, and shared cultural scripts to resolve it as humorous. Monomodal incongruities were rare in this dataset, suggesting that, in these examples, humour was typically achieved through the interplay of modes rather than through a single mode acting alone. While the limited sample prevents broader generalisations, this

tendency highlights how even within very short clips, creators often mobilise multiple semiotic resources in tandem, making the multimodal orchestration of incongruity a particularly salient feature of the humorous effect.

The segmentation approach was central to revealing how these incongruities function. By locating each violation within the sequence of frame-building and frame-breaking moments, it became possible to see whether humour was constructed through gradual layering or introduced immediately, as in remix formats. This temporal mapping also showed that leaving an incongruity unresolved often positioned the viewer's act of reinterpretation as part of the joke. Such insights would not emerge as clearly from analysing moments of tension in isolation, since the timing and relation to the preceding frame are key to how the humour is perceived.

Overall, the findings highlight a genre-specific strategy in which linguistic framing establishes an initial reading that is later playfully violated through embodied or visual modes. Understanding this modal interplay offers a clearer view of how humour in participatory digital media depends on both the precise coordination of modes and the inferential engagement of audiences. Future work could investigate whether similar strategies operate in other TikTok genres or how audiences from different cultural contexts recognise and resolve these incongruities, further clarifying the social and cognitive processes that make them effective.

Bionote: Audrey Willoughby is a doctoral candidate in Linguistics at the University of Milan, in the Department of Languages, Literatures, Cultures and Mediations. She earned her Master's Degree in Applied Linguistics from Texas A&M University-Commerce, where she wrote her thesis *Humor Markers in Computer-Mediated Communication*, and completed a Graduate Certificate in TESOL. Her doctoral research examines how humour is realised semiotically in short-form video platforms, drawing on multimodal discourse analysis to examine meaning-making in digital performance media.

Author's address: audrey.willoughby@unimi.it

References

Abdel-Raheem A. 2018, *Multimodal Humour: Integrating Blending Model, Relevance Theory, and Incongruity Theory*, in "Multimodal Communication" 7 [1], pp. 1-25.

Aiello G. and Parry K. 2020, *Visual Communication: Understanding Images in Media Culture*, SAGE Publications, London.

Attardo S. 1994, *Linguistic Theories of Humor*, Mouton de Gruyter, Berlin.

Attardo S. and Chabanne J.C. 1992, *Jokes as a Text Type*, in "Humor" 5 [1/2], pp. 165-176.

Attardo S. 2020, *The Linguistics of Humor: An Introduction*, Oxford University Press, Oxford.

Attardo S. 2001, *Humorous Texts: A Semantic and Pragmatic Analysis*, Mouton de Gruyter, Berlin.

Attardo S. 2023, *Humor 2.0: How the Internet Changed Humor*, Anthem Press, London.

Attardo S. 2025, *The Shifting Semantics of Point of View (POV)*, in "Salvatore Attardo Blog", https://salvatoreattardo.substack.com/p/the-shifting-semantics-of-point-of (19.9.2025).

Balirano G. and Corduas M. 2008, *Detecting Semiotically-Expressed Humor in Diasporic TV Productions*, in "Humor" 21 [3], pp. 227-251.

Bateman J.A., Wildfeuer J. and Hiippala T. 2017, *Multimodality: Foundations, Research and Analysis – A Problem-Oriented Introduction*, De Gruyter Mouton, Berlin.

Bernad-Mechó E. and Girón-García C. 2023, *A Multimodal Analysis of Humour as an Engagement Strategy in YouTube Research Dissemination Videos*, in "The European Journal of Humour Research" 11 [1], pp. 46-66.

Bucher T. and Helmond A. 2018, *The Affordances of Social Media Platforms*, in Burgess J., Marwick A. and Poell T. (ed.), *The SAGE Handbook of Social Media*, SAGE Publications, London, pp. 233-253.

Gerrig R.J. 1984, *On the Pretense Theory of Irony*, in "Journal of Experimental Psychology: General" 113 [1], pp. 121-126.

Darvin R. 2022, *Design, Resistance and the Performance of Identity on TikTok*, in "Discourse, Context & Media" 46 [100591], pp. 1-11.

Divon T. and Eriksson Krutrök M. 2024, *Playful Trauma: TikTok Creators and the Use of the Platformed Body in Times of War*, in "Social Media + Society" 10 [3], pp. 1-15.

Divon T. and Ebbrecht-Hartmann T. 2022, *Performing Death and Trauma? Participatory Mem(e)ory and the Holocaust in TikTok #POVChallenges*, paper presented at AoIR 2022: *The 23rd Annual Conference of the Association of Internet Researchers*, Dublin, Ireland, November 2022.

Dynel M. 2016, *"I has seen image macros!": Advice Animal Memes as Visual-Verbal Jokes*, in "International Journal of Communication" 10, pp. 660-688.

Gumperz J. 1982, *Discourse Strategies*, Cambridge University Press, Cambridge.

Holsanova J. 2014, *Reception of Multimodality: Applying Eye Tracking Methodology in Multimodal Research*, in Jewitt C. (ed.), *The Routledge Handbook of Multimodal Analysis*, 2nd ed., Routledge, London, pp. 285-296.

Jewitt C. 2014, *An Introduction to Multimodality*, in Jewitt C. (ed.), *The Routledge Handbook of Multimodal Analysis*, 2nd ed., Routledge, London, pp. 15-30.

Kress G. and van Leeuwen T. 2001, *Reading Images: The Grammar of Visual Design*, Routledge, London.

Kułaga W. 2024, *Revolutionizing Visual Communication and Digital Creative Engagement: The Game-Changing Impact of TikTok*, in "Przegląd Socjologii Jakościowej" 20 [3], pp. 212-235.
Lotman Y.M. 1975, *On the Metalanguage of a Typological Description of Culture*, in "Semiotica" 15 [2], pp. 97-123.
Masi S. 2023, *Humour in TED Talks: A Multimodal Account*, in "ESP Today" 11 [2], pp. 328-348.
Raskin V. 1985, *Semantic Mechanisms of Humor*, Reidel, Dordrecht.
Shifman L. 2013, *Memes in Digital Culture*, MIT Press, Cambridge (MA).
Shifman L. 2014, *The Cultural Logic of Photo-Based Meme Genres*, in "Journal of Visual Culture" 13 [3], pp. 340-358.
Trillò T., 2024, *"PoV: You are Reading an Academic Article". The Memetic Performance of Affiliation in TikTok's Platform Vernacular*, in "New Media & Society" 0 [0], pp. 1-22.
Tsakona V. 2020, *Recontextualizing Humor: Rethinking the Analysis and Teaching of Humor*, De Gruyter Mouton, Berlin.
Vásquez C. 2019, *Language, Creativity and Humour Online*, Routledge, London.
Vásquez C. and Aslan E. 2021, *"Cats be outside, how about me?": The Linguistic Construction of Stance in Digitally Mediated Humorous Narratives*, in "Journal of Pragmatics" 171, pp. 240-253.
Wildfeuer J. 2014, *Film Discourse Interpretation: Towards a New Paradigm for Multimodal Film Analysis*, Routledge, London.
Yus F. 2018, *Multimodality in Memes: A Cyberpragmatic Approach*, in Bou Franch P. and Garcés Conejos Blitvich P. (ed.), *Analyzing Digital Discourse: New Insights and Future Directions*, Palgrave Macmillan, London, pp. 105-131.
Zabalbeascoa P. and Attardo S. 2023, *Humour Translation Theories and Strategies*, in L. Kostopoulou and V. Misiou (ed.), *Transmedial Perspectives on Humour and Translation: From Page to Screen to Stage*, Routledge, London/New York, pp. 13-32.
Zacks J.M. and Swallow K.M. 2007, *Event Segmentation*, in "Current Directions in Psychological Science" 16 [2], pp. 80-84.
Zeng J. and Abidin C. 2021, "*#OkBoomer, Time to Meet the Zoomers": Studying the Memefication of Intergenerational Politics on TikTok*, in "Information, Communication & Society" 24, pp. 2459-2481.
Zulli D. and Zulli D.J. 2020, *Extending the Internet Meme: Conceptualizing Technological Mimesis and Imitation Publics on TikTok*, in "New Media & Society" 24 [3], pp. 614-636.