Intermodality:
Combining Linguistic, Audio, and Visual Content for Insight

Jonah Berger
Skyler Chen
Oded Netzer

Jonah Berger (jberger@wharton.upenn.edu) is an associate professor of marketing at the Wharton School, University of Pennsylvania, 3730 Walnut St., Philadelphia, PA 19104. Skyler Chen (skylerymchen@haas.berkeley.edu) is a PhD candidate at the Haas School of Business, University of California, Berkeley, 2220 Piedmont Ave, Berkeley, CA 94720. Oded Netzer (onetzer@gsb.columbia.edu) is the Vice Dean of Research and the Arthur J. Samberg Professor of Business at Columbia Business School, Columbia University, 665 W 130 St. Kravis Hall 941, New York, NY 10027-6902.

Acknowledgements

The authors thank Grant Packard and Giovanni Luca Cascio Rizzo for helpful feedback on the manuscript.

Abstract

From advertisements and social media to sales pitches and customer service calls, marketplace interactions often involve a mix of linguistic (i.e., words), auditory (e.g., sounds), and visual (e.g., images) content. But while these different modalities, or streams of information, frequently interact, most existing research (and practice) analyzes only one modality at a time. Further, while some work has started to consider multiple modalities, it often treats them as independent, failing to consider the interrelationships between them. To fill this gap, the current work introduces the concept of *intermodality*, proposes an integrative theoretical framework regarding different ways modalities relate to, and interact with, each other, and outlines a unifying approach for modeling multimodal communications. This includes (i) systematically aligning levels of representation across modalities, (ii) considering how modalities can be combined (i.e., through separate or joint extraction), and (iii) functional forms and considerations for modeling different types of intermodal relationships. Overall, this work provides insight into how linguistic, audio, and visual content interact, sheds light on how to study intermodality, and offers researchers and practitioners concrete guidance on how to better identify key drivers of relevant outcomes.

Keywords: Language, Audio, Images, Video, Content Analytics, Text Analysis, Unstructured Data, Marketing Communications, Multimodality, Intermodality

Marketplace interactions often involve multiple modalities. Customer service calls include words and audio features (e.g., pitch and tone), face-to-face interactions also contain body language (e.g., gestures), and ads often involve a mix of linguistic (i.e., words), auditory (e.g., sounds), and visual (e.g., images) information.

Further, rather than being independent, these different modalities, or streams of information, often interact (Pieters and Wedel 2004). Consumers are less likely to believe an ad claiming a product is "premium," for example, if it uses a cheap looking image, and vocal tones or facial expressions can make seemingly positive statements (e.g., "I love that brand") seem sarcastic. Consequently, considering multiple modalities, and their potential interactions, leads to better inferences about consumer behavior and helps marketers increase strategic effectiveness (Grewal, Gupta, and Hamilton 2021).

But despite the prevalence of multimodal interactions, most existing research (and practice) analyzes only one modality at a time. Marketing research leveraging unstructured data has primarily focused on text (see Berger et al. 2020 for a review), for example, and while recent research has started to explore audio and visual features (e.g., Fong, Kumar, and Sudhir 2025; Matz et al. 2019; Zhang et al. 2025), it tends to consider individual modalities in isolation. A few papers have started to simultaneously consider multiple modalities (i.e., multimodality, Chang, Mukherjee, and Chattopadhyay 2023; Hartmann et al. 2021; Xu et al. 2025), but this line of research remains limited and often fails to consider *intermodality*, or how content from different modalities *interacts* to shape relevant outcomes.

This paper lays the groundwork to begin to fill these gaps. Specifically, it introduces the concept of intermodality, proposes a framework for different types, and highlights implications for hypothesis generation, data collection, and modeling.

The paper makes three main contributions. First, from a theoretical standpoint, we provide insight into how modalities relate or interact. While individual papers have begun to find cross-modal interactions, there has been little structure around how these disparate findings relate. To address this, we develop an integrative theoretical framework that describes three ways different modalities relate to, and interact with, one another (i.e., content relatedness, interactive effects, and nature of attention). Along the way, we discuss *when* different types of intermodality may be most relevant, and implications for research and practice.

Second, from a methodological perspective, we propose a unifying framework for modeling multimodal communications. This includes (i) systematically aligning levels of representation across modalities, (ii) considering how modalities can be combined (i.e., through separate (late fusion) or joint extraction (early fusion)), and (iii) functional forms and considerations for modeling different types of intermodal relationships.

Third, this work has clear practical implications for a range of marketplace actors. Brands want more persuasive advertisements, call centers want more satisfying customer service interactions, and influencers want to increase engagement. By better understanding how linguistic, auditory, and visual content interact to shape outcomes, companies, organizations, and individuals can increase their impact.

The rest of the paper is organized as follows: We start by defining the core modalities and how they have been studied in marketing, both individually and jointly. Then, we introduce intermodality, and our integrative framework for it, and explore how it should be modeled (i.e., how modalities can be represented and combined). Next, we discuss when modalities should be combined, and which features should be examined in a given context. We conclude with theoretical and practical implications, and directions for future research.

Modalities

While different authors and disciplines use the term modality in different ways, our conceptualization follows the computer science and machine learning literature (e.g., Al-Zoghby et al. 2025; Baltrusaitis, Ahuja, and Morency 2018; Morency and Baltrusaitis 2017), which focuses on three broad types of content: linguistic, visual, and auditory (see Table 1).[1] Just as pipes carry water, and power lines carry electricity, these three modalities carry information.

Table 1: Examples of Linguistic, Visual, and Auditory content

	Linguistic	Visual	Auditory
Example unit	Word	Image	Sound
Multiple units over time	Sequence of words (e.g., sentence)	Sequence of images (e.g., video)	Sequence of sounds (e.g., audio recording)

Linguistic content includes elements like words, sentences, and longer expressions, and can be conveyed through written or spoken means.[2] Visual content includes any non-linguistic information that comes in through the eyes, such as images, or sequences of them (e.g., video).[3] Auditory content includes any non-linguistic information that comes in through the ears, such as vocal cues, music, and other sounds.

Different data sources can contain different modalities. Audio recordings, for example, may contain just auditory content (i.e., sounds), or a mix of audio and linguistic content (e.g., a conversation). Videos always contain visual content (i.e., a sequence of images) but can also

[1] While smell, taste, and touch also carry information, given the relative lack of tools and data that allow these modalities to be examined, we do not focus on them here, and discuss them further in the General Discussion.
[2]Given linguistic content is quite different than other sounds or images, following work in communications (New London Group 1996), we treat it as a separate modality.
[3] Some work in communications and semiotics (e.g., New London Group 1996) further separates visual signals into aspects like body language and other types of visual communication (e.g., charts) but given that empirical work has not separated the two, we treat them all as visual content and discuss separating them in the General Discussion.

contain linguistic and auditory information. Throughout the paper, we use the term modalities to describe the different types of information streams and distinguish them from the data sources (or media; Grewal, Gupta, and Hamilton 2021) used to study those streams.

Modalities in Marketing

As discussed, marketplace actors often consume (and produce) content from multiple modalities. Restaurant reviews may use language to describe the service, for example, and images to show how delicious the food was. Salespeople may use language to explain why someone should buy, while using vocal and visual cues to signal they can be trusted.[4]

Consequently, audiences often use information from multiple modalities to decode communications (e.g., using both the text and images of reviews to pick a restaurant). As a result, responses are rarely driven by one modality alone, and emerge from the combined, interacting signals embedded across formats. Understanding such intermodality therefore requires not only combining modalities (i.e., multimodality) but examining how modalities reinforce, contradict, or transform one another to drive responses.

Most marketing research analyzing unstructured data, however, is unimodal (i.e., examines only one modality), and the great majority has focused on language. Given the rise of user-generated content (e.g., consumer reviews and social media posts), for example, and the fact that it is often readily accessible for analysis, a burgeoning stream of literature has used automated textual analysis to study language (see Berger et al. 2020; Hartmann and Netzer 2023; Humphreys and Wang 2017 for reviews). Many marketplace interactions (e.g., customer service

[4] Some situations are inherently more multimodal than others. While written language can occur by itself (e.g., in a text), for example, or with other modalities (i.e., if the text contains an image), spoken language is always accompanied by audio features (e.g., pitch). Similarly, videos always contain a sequence of images, though they often also contain language and audio features.

calls and television advertisements) also contain audio features (e.g., pitch and tone) and advances in automated audio analytics have encouraged work in this area (see Hildebrand et al. 2020; Wang, Bendle, and Pan 2024 for reviews). And bolstered by recent advances in computer vision, the popularity of image-based social media platforms (e.g., Instagram), and wider availability of video data, research has started to study how visual features shape perceptions and behavior (see Dzyabura and Peres 2021; Villarroel Ordenes and Zhang 2019; Wang, Bendle, and Pan 2024 for reviews). See Web Appendix for a detailed discussion of unimodal work.

Work that does examine multiple modalities (i.e., multimodality) has mostly focused on what can be extracted from each, rather than meaningfully assessing interactions among them. Some papers treat one modality as focal while including others as separate, independent controls (e.g., Cascio Rizzo and Berger 2023; Chang, Mukherjee, and Chattopadhyay 2023; Kim, Jiang, and Thomadsen 2023). Others include two or more modalities to examine what each conveys about the outcome of interest, but do not model interactions among them (e.g., Balducci et al. 2026; Boughanmi and Ansari 2021; Chakraborty et al. 2025). While these approaches may be appropriate for certain research questions, they ignore potentially meaningful interactions.

A small set of papers have started to explicitly consider interactions among modalities (e.g., Cascio Rizzo, Berger, and Zhou 2025; Ceylan, Diehl, and Proserpio 2024; Villarroel Ordenes et al. 2019). Yazdani, Chakravarty, and Inman (2025), for example, extract emotions from images and text to examine how their (mis)alignment shapes donor engagement. But while each paper uncovered useful findings, because they focused on a single methodological approach and empirical setting, they have less to say about the broader nature of intermodality.

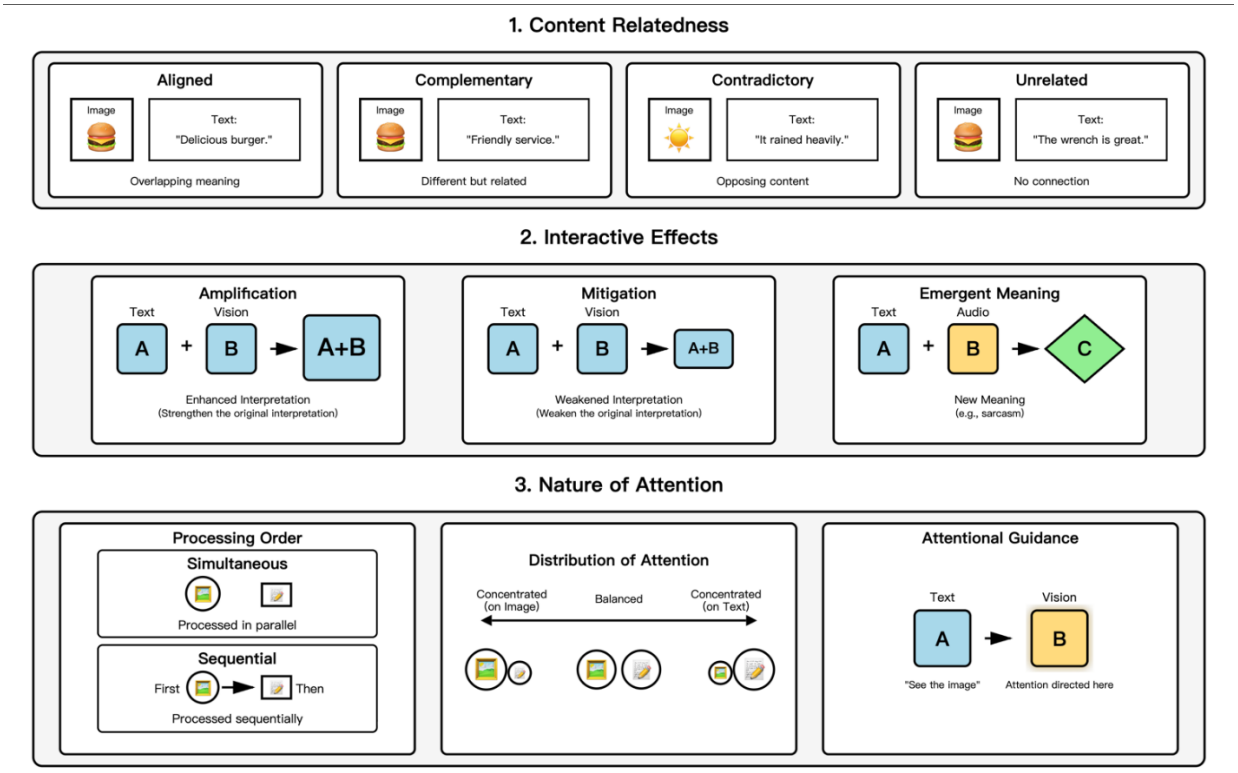
Building on this work, we propose a structured framework outlining different types of intermodality and discuss methodological considerations involved in modeling them.

Types of Intermodality

Drawing on literature in psychology, communication, marketing, and other disciplines, we propose an integrative framework that sheds light on how multiple modalities relate or interact. While individual papers have identified specific cross-modal phenomena, from perceptual fusion in sensory research (Alais and Burr 2004; McGurk and MacDonald 1976) to congruence effects in advertising (Heckler and Childers 1992), there has been little integration.

We organize disparate findings around three theoretically distinct dimensions: *content relatedness* (i.e., the relationship between content streams), *interactive effects* (i.e., whether combining modalities produces outcomes that exceed, fall short of, or fundamentally differ from what each would produce in isolation), and the *nature of attention* (see Figure 1 and Table 2).

Figure 1: Types of Intermodality



Note: For squares with letters, color conveys meaning or outcome, and size reflects strength or intensity.

Content Relatedness

Content relatedness describes how content from different modalities relates to one another. Communication theory has long recognized that messages can be simultaneously carried by multiple channels, and that the correspondence between them shapes how receivers decode meaning (Kress and Van Leeuwen 2001). Schema and congruence theories further suggest that the structural match between incoming cues shapes processing fluency and evaluation (Heckler and Childers 1992; Meyers-Levy and Tybout 1989). Content relatedness can thus be a useful first step in assessing interrelationships among modalities, and can occur along multiple dimensions (e.g., semantic or emotional relatedness).

More specifically, modalities can (a) convey similar things (i.e., be *aligned*), (b) address related but distinct topics (i.e., be *complementary*), (c) carry conflicting signals (i.e., be *contradictory*), or (d) bear no meaningful relationship (i.e., be *unrelated*).

Aligned

When content from different modalities is *aligned*, they convey similar meanings. When an online review talking about a hamburger also contains a picture of a burger, for example, the semantic alignment provides mutually reinforcing information. The same holds when a movie's soundtrack and visuals evoke similar emotions (i.e., emotional alignment), or when a tweet's language and image involve similar speech acts (i.e., communicative function alignment).

Such alignment should facilitate a coherent impression, reinforce interpretation, enrich understanding, and make content easier to process.[5] Indeed, reviews are more helpful when the text and photos are semantically similar or congruent (Cao, Li, and Zhang 2025; Ceylan, Diehl,

[5] If the information from multiple channels is too redundant, though, it may sometimes backfire.

and Proserpio 2024), music NFTs are valued more when audio and album art are emotionally similar (Ding and Wang 2023), and presentations are more persuasive when gestures are aligned with verbal content (Cascio Rizzo, Berger, and Zhou 2025).

Complementary

Even if they don't convey similar meanings, content from different modalities can address *complementary* or related topics or themes. Similar to when content is aligned, such *complementarity* should enrich interpretation. Semantic complementarity (e.g., when a review of a restaurant's friendly service includes a picture of a hamburger), for example, can offer a richer understanding of the situation (i.e., the meal) than either modality alone. Complementarity may also have other benefits. Complementary communicative intent between images and text, for example, may increase novelty, which increases social sharing (Villarroel Ordenes et al. 2019).

Complementarity may be particularly valuable when each modality captures content that is difficult to convey through the other (something that we discuss later). It's difficult to fully capture the quality of a restaurant's service in a picture, for example, or its décor in words.

Contradictory

Content from different modalities can also be *contradictory*. When an influencer says "I'm excited to partner with this brand" in a flat and disengaged tone, for example, the audio conflicts with the linguistic content. The same goes for happy imagery combined with a sad audio track (i.e., emotional contradiction).

Such contradiction can generate skepticism and reduce perceived credibility (Gillis and Nilsen 2017). That said, some contradiction can evoke curiosity or arousal (Berlyne 1960) and

increase processing as people try to reconcile it (Meyers-Levy and Tybout 1989). This may have benefits, such as increased clicks, memorability, and even purchase likelihood in some instances (Cao, Li, and Zhang 2025; Heckler and Childers 1992).

Unrelated

Finally, though it occurs less frequently, content from different modalities can be unrelated (i.e., not connected in any way). Content producers usually want communication to be coherent, but unrelatedness may arise accidentally (e.g., automatically generated thumbnails that don't fit the caption), or due to human error. Alternatively, it may occur strategically, when creators use unrelated content to grab attention (e.g., clickbait).[6] The lack of meaningful connection between modalities should make it difficult to form a coherent interpretation, however, which may create confusion and lead to negative downstream consequences (e.g., reviews being less helpful).

Interactive Effects

Separate from whether (or how) content is related, the co-presence of multiple modalities can generate interactive effects that go beyond their independent contributions.[7] A foundational insight from multisensory perception research is that the brain does not simply add signals from different channels; it integrates them in ways that can be supra-additive or sub-additive (Partan and Marler 1999; Welch and Warren 1980). The McGurk effect (McGurk and MacDonald

[6] Content may also be unrelated when it is created by different producers (e.g., display ads visuals and the text of the webpage they appear on). We further discuss intermodality across multiple content pieces in the General Discussion.
[7] While content relatedness may provide insight into how likely different types of interactive effects are to occur, one may also just interact features from different modalities without first examining their relatedness. Further, note that unlike content relatedness, which is independent of the outcome of interest, interactive effects are often outcome dependent (i.e., two modalities may amplify one outcome and mitigate another). Clickbait display ads where the image and text contradict each other, for example, may lead to higher click-through-rates but lower conversion.

1976), for example, finds that when auditory and visual speech cues conflict, the perceptual system fuses them into a new percept that corresponds to neither channel alone. These interactive effects ultimately determine whether multimodal communication achieves its goals: whether an ad is persuasive, a review is credible, or a service interaction is satisfying. Characterizing them requires explicitly modeling how modalities combine, rather than treating each as independent.

Interactive effects include (a) amplification, (b) mitigation, and (c) emergent meaning.

Amplification

In some situations, content from multiple modalities may reinforce each other to produce a greater effect (i.e., supra-additivity, where $f(A,B) > f(A) + f(B)$). An ad describing a dish as "spicy" while showing visuals of chili peppers, for example, amplifies the perceived spiciness. Amplification can arise when modalities are semantically aligned (e.g., Dang et al. 2026), such that consistent meaning across modalities reinforces interpretation, or when modalities contain complementary information, as distinct but mutually reinforcing cues can create a coherent overall theme that strengthens interpretation.

Mitigation

In other cases, however, one modality may reduce the impact of another. When someone says "I'd be happy to" in a flat voice while looking away, their tone and body language undermine the positive language and suggest that they're not really interested. Mitigation can be thought of as sub-additivity such that $f(A,B) < f(A)+ f(B)$, where the joint effect of the modalities in shaping outcomes is weaker than their additive effect. Mitigation can be driven by

contradiction (e.g., Lee 2021), as conflicting cues reduce the meaning conveyed by each, but can also occur with aligned content through a ceiling effect or diminishing returns (concavity).

Emergent meaning

Content from multiple modalities can also generate new or emergent meanings $f(A,B) = f(C)$ (Forceville 1996). When a highway billboard says "hot and fresh, exit here," for example, and shows a picture of fried chicken, the integration of the linguistic and visual cues provides the emergent meaning that there is a restaurant nearby with fresh fried chicken. Neither the linguistic nor the visual content creates that meaning alone, but when combined, that meaning emerges.

Nature of Attention

Finally, beyond how content relates, or potential interactive effects, one can also examine how the co-presence of multiple modalities shapes attention (and its downstream consequences). Because attentional capacity is limited, the brain cannot process all modalities simultaneously with equal depth (Mayer 2001). Dual coding theory (Paivio 1971) proposes that linguistic and nonlinguistic information draw on distinct cognitive subsystems, each with its own processing constraints, which shapes how attention is allocated (and integrated) across modalities. Attention is also dynamic: work on gaze (Rayner et al. 2001) and visual salience (Pieters and Wedel 2004) shows that attention shifts across modalities, and that earlier-attended cues can frame and bias the interpretation of later ones.

Types of intermodality related to the nature of attention include (a) processing order, (b) the distribution of attention, and (c) attentional guidance.

Processing order

Content from different modalities can vary in the order in which they are processed. This influences which modality is salient or diagnostic, and how cues are integrated and interpreted.

Multimodal cues are often processed *simultaneously*. When listening to an in-person sales pitch, for example, listeners are simultaneously considering what is being said (i.e., the words) as well as the salesperson's facial expressions and vocal tone. Consequently, listeners may fuse information across modalities to form a unified impression (e.g., perceiving the salesperson as untrustworthy if their words and tone don't align). Simultaneous presentation can thus lead to more integrated processing (Mayer and Anderson 1992), though it can also divide attention across modalities, leading content from each to receive less attention (Moreno and Mayer 1999).

In other situations, however, content from different modalities is processed more *sequentially*, even when they are part of the same communication (Rayner et al. 2001). When reading newspapers, for example, early attention is often drawn to salient lead photos, with text receiving little attention in the initial seconds (Bucher and Schumacher 2006). Similarly, when reading an email, people usually start with the subject line and only then look at any visual content (e.g., attached images) that may be inside.[8] In such sequential situations, earlier cues may frame or bias the interpretation of later ones. A shampoo bottle surrounded by green leaves, followed by the tagline "Now with new materials," for example, may lead viewers to infer that the product is eco-friendly.

Note that the modality through which information is conveyed may shape the processing order. When language is spoken, for example, it is processed through the auditory channel, and can thus be simultaneously integrated with visual information (e.g., the speaker's facial

[8] The spatial relationship between linguistic and visual content (e.g., whether words are overlaid on top of images, Dang et al. 2026) may shape which modality is processed first.

expressions and vocal tone). When language is written, however, processing shifts to the visual channel, and thus it becomes more difficult to simultaneously process other visual information (e.g., an image; Mayer 2001). Consequently, the temporal consumption of multimodal information may depend, in part, on how language is delivered.

Further, even when multimodal content is *delivered* simultaneously, differences in processing speed may lead certain modalities to be processed more quickly. The gist or basic meaning of images can be grasped in roughly 100ms (Potter 1976), while words usually elicit slower semantic processing (Kutas and Federmeier 2011). Nonlinguistic sounds also tend to be interpreted faster than spoken words (Cummings et al. 2008). Consequently, when someone raises their voice, listeners may sense anger even before they process the words being said.

Distribution of attention

Independent of the order in which modalities are processed, audiences may allocate their attention in different ways. In some situations, attention is distributed rather equally. Listening to a speaker, for example, often involves *balanced* attention to both language and tone of voice or intonation to interpret meaning. In other cases, however, attention is primarily *concentrated* in one modality. When reading online news, for example, language receives most of the attention, while visual content (e.g., visual ads) tends to receive less (Simonov, Valletti, and Veiga 2025).

Which modality receives more attention may be driven by expectations (e.g., Pinterest users know it is an image focused platform) or perceptual salience, as salient cues (e.g., bright colors or loud sounds) naturally draw attention. It may also vary over time (e.g., a speaker raising their voice to highlight certain words). Certain people also rely more on certain modalities (e.g., auditory vs. visual learners; Fleming and Mills 1992) and task- and goal-dependence can also

play a role, with some tasks or goals encouraging attention towards a particular modality (e.g., reading instructions vs. interpreting a diagram).

Attentional guidance

Though less frequent, modalities may also direct attention toward one another. Verbal content (e.g., "see the photo below") can prompt viewers to focus on images, for example, and subtitles (e.g., "listen for the chime") can draw attention to sounds. Such attentional guidance shifts how attention is allocated across modalities, and almost always implies sequential processing, where the guiding modality is processed first.

Table 2: Types of Intermodality

	Sub-dimension	**Definition and Marketing Relevance**
Content Relatedness	Aligned	How content from different modalities relates to one another (i.e., whether it is aligned, complementary, contradictory, or unrelated). A salesperson's gestures, for example, may illustrate what they are saying or be unrelated
	Complementary	
	Contradictory	
	Unrelated	
Interactive Effects	Amplification	The co-presence of multiple modalities generates interactive effects (i.e., supra-additivity, sub-additivity, or new meaning) that go beyond their independent contribution. The combination of words and images in an ad, for example, can create different meanings than either modality separately.
	Mitigation	
	Emergent Meaning	
Nature of Attention	Processing Order	The co-presence of multiple modalities shapes the order and distribution of attention. In print ads, for example, visual content captures initial attention, while text size drives sustained attention (Pieters and Wedel 2004).
	Distribution of Attention	
	Attentional Guidance	

Modeling Intermodality

To understand reactions to intermodal content, predict its effectiveness, and design more effective communications, intermodality needs to be modeled in a way that mimics, or captures how audiences perceive and integrate content.

Linguistic, visual, and auditory content are all forms of unstructured data, which raises several modeling challenges. First, each modality is often high-volume and high-dimensional and combining modalities quickly compounds the problem (Grimmer, Roberts, and Stewart 2021). Second, unlike structured data, unstructured inputs generally require feature extraction or representation to produce meaningful insights or structured information for further analysis. These transformations entail information loss in ways that may differ across modalities (Bengio, Courville, and Vincent 2013) and make it difficult to relate them on a common scale. Third, modalities encode information in fundamentally different forms. Linguistic data are discrete and compositional, carrying semantic, pragmatic, and syntactic relations (Hockett 1960; Jackendoff 2002). Visual data are spatial, reflecting low-level properties (e.g., color), mid-level design attributes (e.g., symmetry), and high-level content (e.g., faces). Because encodings are heterogeneous, integrating modalities to capture intermodality is not straightforward and requires fusion approaches that enable meaningful joint inference (Al-Zoghby et al. 2025).

Each modality can be thought of as a noisy measurement of latent constructs (e.g., sentiment or product features). To be able to extract them, we need to convert the unstructured data into the latent constructs. Thus, we first discuss how individual modalities can be represented, and then turn to the joint modeling of multimodalities.

Levels of Representation

What can be extracted from a modality depends on how the content is represented. Different levels of representation (i.e., simple features or high-dimensional embeddings) extract different amounts and types of signals, which have implications for how to capture and model cross-modal relationships (see Table 3 for tools, techniques, and applications in marketing).

Table 3: Different Representations Across Modalities

	Language	Audio	Image	Video
Simple Representations	• **Dictionaries** - LIWC (Berger and Milkman 2012; Netzer, Lemaire, and Herzenstein 2019); Certainty Lexicon (Rocklage, Rucker, and Nordgren 2018) • **Part-of-speech tagging** (Netzer et al. 2012; Packard and Berger 2020) - Natural Language Toolkit in Python; tm in R • **N-grams** (Netzer et al 2019; Rocklage et al. 2023) - NLTK in Python; Ngram in R	• **Tools** - openSMILE; Librosa; Echo Nest; Pratt • **Loudness** (Cascio Rizzo and Berger 2023; Xu et al. 2025) • **Brightness (**Xu et al. 2025) • **Pitch (**Cascio Rizzo and Berger 2023) • **Tempo** (Yang et al. 2022) • **Timbre** (Efthymiou et al. 2023) • **Tone** (Wang et al. 2021)	• **Tools** - skimage.feature module in scikit-image; OpenCV feature extraction • **Color** (Cao, Li, and Zhang 2025; Dew, Ansari, and Toubia 2022; Li and Xie 2020) • **Color contrast** (Cao, Li, and Zhang 2025) • **Symmetry (**Dew, Ansari, and Toubia 2022) • **Edges** (Dew, Ansari, and Toubia 2022)	Representation of frames of video is similar to images, but the temporal nature of video also offers motion features. • **Motion magnitude, motion direction, and foreground motion area** (Zhou et al. 2021) • **Scene cuts** (Liu et al. 2018)
Intermediate Features	• **Topic modeling** (Berger and Packard 2018; Li and Ma 2020; Netzer, Lemaire, and Herzenstein 2019; Tirunillai and Tellis 2014) - LDA; LSA	• **Spectrogram (**Cascio Rizzo and Berger 2023) - Librosa in Python • **Mel-spectrogram** (Fong, Kumar, and Sudhir 2025) - Librosa in Python • **Turn-taking dynamics** (Balducci et al. 2026; Cascio Rizzo and Ziano 2025)	• **Face detection** (Feng et al. 2025) – OpenCV, Dlib • **Semantic label detection** (Ceylan, Diehl, and Proserpio 2024) - Google Cloud Vision API • **Human keypoint detection** (Cascio Rizzo, Berger, and Zhou 2025) – MediaPipe • **Facial expressions** (Cascio Rizzo, Berger, and Villarroel Ordenes 2023; Li and Xie 2020) - Google Cloud Vision API; OpenCV • **Gaze estimation, face landmark detection and tracking** - OpenFace	Representation of frames of video is similar to that of images, but the temporal nature of video also offers motion features. • **Optical flow** (Zhou et al. 2021) - OpenCV
Richer Representations	• **Embeddings** - Word2Vec (Timoshenko and Hauser 2019; Toubia, Berger, and Eliashberg 2021); GloVe (Hong and Hoban 2022) • **Transformer-based** (Cao, Li, and Zhang 2025) - BERT (Hartmann et al. 2023) • **AI tools** (Arora, Chakraborty, and Nishimura 2025; Blanchard et al. 2025; Rathje et al. 2024) - OpenAI GPT; Google Gemini; Anthropic Claude.	• **Embeddings (CNN-based)** - VGGish; OpenL3 • **Transformer-based -** wav2vec • **AI tools** - OpenAI Whisper API; Google Cloud Speech-to-Text; Microsoft Azure Speech	• **Embeddings (CNN-based)** - ResNet (Cao, Li, and Zhang 2025; Feng, Li, and Zhang 2025); CLIP (Hartmann, Exner, and Domdey 2025) • **Transformer-based** - Vision Transformer (Zhu et al. 2025) • **AI tools** (Exner et al. 2025; Huang and Katona 2025; Nanne et al. 2020) - OpenAI Vision; Google Gemini Vision; AWS Rekognition; Microsoft Azure Vision.	• **Transformer-based** (Bravin et al. 2025) - ViT • **AI tools** - OpenAI GPT video input; Google Gemini Flash; Meta Llama vision; RunwayML; AWS Rekognition.

Simple representations often focus on low-dimensional features. These representations are easier to interpret, but tend to be limited in the aspects of the information conveyed. In language, for example, dictionaries (e.g., Boyd et al. 2022) are often used to capture psychologically meaningful categories (e.g., number of positive words), and part-of-speech tagging helps analyze grammatical components of texts. In audio, acoustic properties, such as loudness, brightness, and speech rate can be extracted using toolkits like Librosa. And for visual content, simple image representations include color, hue, and symmetry (extracted using tools like OpenCV) with motion-based features like magnitude and direction added for video.

Intermediate representations are more structured features that summarize patterns, categories, or relationships, while remaining more interpretable than richer embeddings. In language, topic modeling techniques (e.g., Latent Dirichlet Allocation (LDA)) can uncover underlying themes. In audio, raw sound waves can be converted into numeric representations (e.g., Mel-spectrograms) that summarize spectral patterns over time and can be visualized as images. In visual content, intermediate features include face detection, facial expressions, and hand gestures, with video adding temporal features like optical flow.

Richer representations are high-dimensional, learned embeddings that can capture more abstract structure and complex relationships. In language, models have evolved from Word2Vec and GloVe to contextual transformer-based approaches like BERT, and more recently to LLMs capable of extracting nuanced information from text (Blanchard et al. 2025; Rathje et al. 2024). Similar embedding approaches exist for audio (e.g., VGGish) and visual content (e.g., ResNet). Generative AI tools have become particularly valuable for image and video analysis. Rather than tagging discrete features, they produce holistic, narrative-style descriptions that more closely mirror how humans interpret content. While an intermediate representation model may extract

features like [woman], [baby], [smile] from an image of a mother hugging her newborn baby, for example, generative AI is likely to analyze the image as "a mother hugging her newborn baby."

Overall, different levels of representation differ in interpretability and the amount of relevant information provided. Simple representations capture specific constructs or features (e.g., linguistic certainty or auditory loudness), making them interpretable, but less capable of capturing the multitude of information contained. Intermediate representations can summarize features into something more structured (e.g., topics in text or facial expressions in images), while remaining interpretable. Richer representations are least interpretable, but most holistic, and tend to maximize signal-to-noise. Though they may not be easily mapped onto certain constructs, they provide a more comprehensive view of the data, which is useful for tasks like similarity comparisons, clustering, or predictive modeling, where the full richness is essential.

Combining Modalities - Early vs. Late Fusion

Beyond how each modality is represented, though, it's also important to consider how to model potential intermodal interactions, and whether to extract information from each modality separately (late fusion) or jointly (early fusion).

Separate extraction

Most multimodal marketing work uses separate extraction (i.e., *late fusion*). These approaches first extract, or represent separately, the information from each modality (at any level of representation) and then combine them (with or without accounting for interactions) in subsequent analyses. To compute the semantic similarity of the photos and text of online reviews, for example, Ceylan, Diehl, and Proserpio (2024) first used Google Cloud Platform

Vision API to extract image labels and Doc2Vec to create vectors for reviews and photo labels.

This approach provides flexibility. One can use the most appropriate level of representation for each modality without worrying about how to combine them into a common space (e.g., concatenating visual features like color with audio features like pitch). It also offers simplicity, using off-the-shelf tools to extract information from each modality. Because information from each modality is extracted separately, though, there is no guarantee that it shares commonalities, which limits the ability to capture cross-modal interactions.

Joint extraction

To capture rich and complex interactions among modalities, it is often useful to extract information jointly (i.e., *early fusion*), and prior to relating that joint multimodal information to the outcome variable. Cross-modal attention is a common approach. In cross-attention transformers, queries from one modality (e.g., linguistic content) may attend to the keys and values from another (e.g., visual content), allowing the model to learn how the features in one modality relate to those in another.

Several models and tools allow for joint extraction. Contrastive Language–Image Pretraining (CLIP), for example, uses an image and a language encoder, to map their respective inputs into a vector representation of the same dimensionality, providing a natural way for the two modalities to interact. Multimodal graph neural networks, where each modality can be represented as a set of interconnected nodes and edges, enable the model to dynamically update feature representations based on relationships learned during training. Luo et al. (2025), for example, used a graph memory fusion network (GMFN) to simultaneously fuse linguistic, auditory, and visual content, as well as their interactions. Other architectures, such as structured

multimodal variational autoencoders (VAE), learn interactions implicitly by first training modality-specific encoders, and then combining them in a joint neural network that allows for flexible, non-linear interactions between modalities before projecting to a common latent space (Dew, Ansari, and Toubia 2022; Tian, Dew, and Iyengar 2024). This architecture allows dependencies between modalities to be captured during representation learning.

Generative AI tools (e.g., GPT-4 Vision or Google Gemini) are also well-suited to jointly extract information across modalities. They convert modalities into a shared token representation and process them in a unified transformer architecture that learns cross-modal relationships through attention. These models are flexible enough to perform different types of tasks, such as understanding texts, classifying objects in a scene, and transcribing and analyzing audio (Blanchard et al. 2025). That said, as is evident by their name, LLMs are inherently language-based. Consequently, all modalities are eventually translated into linguistic information. While this is appealing for multimodal analysis, it means that these models require an additional "translation" for non-linguistic modalities, which may lead to some information to be lost.

Hybrid fusion

While joint extraction is beneficial because it allows for rich and complex interactions across modalities, the need to project all modalities onto the same space can be limiting when features don't share common space (e.g., color in image data versus pitch in audio data). To address this limitation, some researchers use *hybrid fusion,* which combines joint extraction of information with independent extraction of features from each modality (Al-Zoghby et al. 2025).

Joint extraction approaches also often lack clear interpretability. Because interactions are learned implicitly in high-dimensional latent spaces, they are not directly interpretable in the way

post-extraction interactions are. Tools like SHAP (SHapley Additive exPlanations), or attention weight analysis, can provide approximations for modality contributions. Such tools help address the tradeoff between predictive performance and interpretability when choosing between interaction-during-training models and post-extraction interaction.

Modeling Different Forms of Intermodality

Importantly, however, what levels of representation to use, and how to model cross-modal interactions, depends in part on what type of intermodality is being considered (Table 4).

Content relatedness

Content relatedness (e.g., alignment or unrelatedness) is focused on the way the different modalities relate to one another (rather than an outcome). Consequently, modeling decisions are focused on the relationship between extracted features (rather than their interactive effect). This often involves encoding all modalities into a shared (e.g., joint embedding) space and computing a similarity metric (e.g., cosine similarity) between the resulting representations. While features of each modality could be extracted separately and then related, an early fusion approach allows the content from different modalities to relate in a more complex and nonlinear manner. Models pre-trained on paired multimodal data (e.g., CLIP for image-text pairs) are particularly well suited here, as their training objective explicitly aligns representations across modalities.

Specific types of content relatedness should also be examined differently. *Alignment* or *unrelatedness*, for example, can be captured by computing the similarity between each modality's representation, where high scores imply alignment, and zero suggests unrelated. A restaurant review describing the food should have a positive similarity score if the accompanying

image nicely represents the dish, for example, but a score closer to zero if it shows the restaurant's decor. One should be cautious, however, about assuming similarity indicates alignment without validating that the embedding space is sensitive to the specific dimension of interest (e.g., emotional vs. topical similarity). To address this, one can separately examine the similarity of lower-level representations relating to content (e.g., topics) and emotion.

Table 4: Modeling Different Forms of Intermodality

	Fusion Approach	Representation / Feature Extraction	Modeling Approach	Key Considerations
Content Relatedness				
Alignment	Early or Late	Joint embedding space (e.g., CLIP).	Cosine similarity between modal representations; high score = alignment, near zero = unrelatedness.	May need separate topical versus emotional alignment.
Contradiction	Early or Late	Content and valence extraction per modality.	Directional valence similarity score; negative = contradiction, positive = alignment.	Standard similarity metrics may conflate contradiction with unrelatedness; directional disagreement must be modeled explicitly.
Complementarity	Early (preferred)	Separate topic models (e.g., LDA or BERTopic) per modality; matrix factorization.	Relatedness term (embedding similarity) + redundancy penalty (mutual information). Matrix factorization can identify co-occurring themes that jointly drive outcomes.	Can be complex to operationalize, requires capturing thematic overlap alongside informational distinctiveness.
Unrelatedness	Late	Modality-independent feature extraction.	Test for relatedness before including interactions.	Near zero similarity may signal unrelatedness or two opposing forces. Need to examine different dimensions of similarity.
Interactive Effects				
Amplification	Early or Late	Modal signals extracted independently; no shared space required.	Multiplicative interaction term.	Standardize modal features before forming product term.
Mitigation	Early or Late	Modal signals extracted independently; no shared space required.	Multiplicative interaction term. Distinguish contradiction-driven mitigation (directional conflict score) from ceiling/diminishing returns (concave transformation).	The two sources of mitigation (contradiction and diminishing returns) have different functional forms and should not be conflated.
Emergent Meaning	Early (required)	Joint encoding across modalities using early fusion.	Can use supervised learning or LLMs for labeling specific emergent meanings, validated by human annotation.	Ground-truth labeling is difficult and human validation is essential.
Nature of Attention				
Processing Order	Early (preferred)	First modality as conditioning context for second.	Early fusion or cross-modal attention allowing earlier modality to attend to later one. In regression: asymmetric interaction term where first modality moderates the second's effect but not vice versa.	Eye tracking can validate assumed order; allow for heterogeneity across users in the order.
Distribution of Attention	Early or Late	Use Platform priors or eye-tracking dwell times per modality.	Attention-weighted fusion with unequal modality weights; priors from platform context could serve as priors. Multilevel models allowing modality weights to vary by person or context.	Eye-tracking dwell times can provide empirical priors for fusion weights
Attentional Guidance	Late	Guidance cue detection based on NLP or visual element recognition; feature re-weighting in guided modality.	Modality-conditioned attention. In regression: binary guidance indicator × guided modality features as interaction term; test whether guidance amplifies the specific targeted feature.	Guidance implies sequential processing; guided modality features should be re-weighted toward the content the cue targets.

When trying to examine *contradiction*, however, standard similarity metrics may conflate it with unrelatedness. Consequently, it may be useful to consider directional valence disagreement, in addition to content similarity. A positive food review accompanied by an image of disgusting-looking food should have moderate or high content similarity (i.e., both involve food), for example, but a negative valence similarity, implying contradiction. Contradiction can also be captured by building an early-fusion (e.g., transformer) model trained to capture contradictory content across modalities.

Complementarity is harder to operationalize because it involves covering related topics or themes in distinct or non-overlapping ways. As a result, it requires capturing both thematic relatedness and informational distinctiveness. One approach may be to decompose this into (1) a relatedness term (operationalized via embedding similarity), and (2) a redundancy penalty (which can be estimated by measuring the mutual information between the two modal representations or by checking whether the topic distributions extracted from each modality share themes). Because of the complexity of this type of cross-modal relationship, early fusion approaches are likely more appropriate. Following Ruiz, Athey, and Blei (2020), complementarity can also be measured using matrix factorization to decompose the content of each modality into its underlying themes (e.g., what topics a review's text covers), and then estimate whether different modalities' themes systematically co-occur in ways that drive outcomes above and beyond what either does alone.

Interactive effects

Interactive effects examine how cross-modal interactions shape outcome variables. This doesn't require capturing modalities at the same level of representation and can use late (rather

than early) fusion methods. *Amplification (mitigation)* implies a supra (sub)-additive functional form: the effect of both modalities together is greater (less) than the sum of their individual effects. A multiplicative interaction term between two modalities can be used to capture such effects. In linear regression contexts, this takes the form $\hat{y} = f(\beta_1 X_1, \beta_2 X_2, \beta_3 X_1 X_2)$ where, for example, for positive *X*s and the positive effect of each modality independently ($\beta_1 > 0$ and $\beta_2 > 0$) a significantly positive β_3 indicates amplification (Ceylan, Diehl, and Proserpio 2024), and a negative one indicates mitigation (Lee 2021).

That said, it is important to distinguish between mitigation driven by contradiction (i.e., subadditivity) versus ceiling effects or diminishing returns. The former calls for an interaction term sensitive to directional disagreement, while the latter calls for a concave functional form (e.g., logarithmic or square-root transformation) of one or both modal variables before combining them, or a saturating activation function in the neural network specification. The two approaches can be combined when both mechanisms may be operating simultaneously.

Emergent meaning is perhaps the most challenging to model because, by definition, it requires processing the two modalities jointly with a specific relationship in mind. Multimodal transformers that process modalities in relation to each other early in the fusion pipeline can enable the model to identify concepts that span both modalities. Because emergent meaning takes the form of a new semantic category or entity (e.g., inferring that a combination of text and image implies a nearby location), supervised learning approaches with ground-truth labeling, often using human annotators, of emergent meanings for supervised training are often needed. LLMs can also be used to explore specific forms of emergent meanings. However, the LLM labeling should also be validated by human labeling.

Nature of attention

The nature of attention to different modalities, also has modeling implications. The discussion so far has been agnostic about *processing order* (i.e., the order in which different modalities are processed), but if clear ordering exists,[9] this can impact how intermodality should be modeled. Sequential exposure means that earlier modalities shape the interpretation (or impact) of later ones, and modeling this requires architectures that encode temporal asymmetry. Recurrent architectures (LSTMs, GRUs) applied across modalities in sequence (Hochreiter and Schmidhuber 1997) could be used, but a cleaner approach is to use the first modality's representation as a conditioning context for encoding the second, implemented via cross-modal attention. In regression models, this can be approximated by including an interaction between the modalities as in moderation effects, but the interpretation is explicitly directional: the first is treated as a "moderator" of the second, but not the other way around.

Attention to different modalities varies not only temporally (i.e., processing order), but also in magnitude (i.e., *distribution of attention*). This has implications for how much weight each modality's signal should receive in a model. While a larger coefficient on the relationship between a specific modality and an outcome may signal attention imbalance, other factors (e.g., differences in the impact of modality on the outcome for a given unit of attention or variation in processing speed between modalities) may also account for such differences.

Eye-tracking data is perhaps the most natural way to directly measure the distribution of visual attention, as it provides a continuous behavioral record. Rather than relying on proxies (e.g., platform norms or self-report), fixation counts, or how long consumers spent examining

[9] Ordering can be inferred from the context (i.e., images are likely to be processed first on image-based platforms like Instagram) or examined empirically (e.g., through eye tracking). Measuring the order, and how long each modality was attended to before the other, can also test the strength of potential ordering effects.

text versus image content, can capture attention. Consequently, it can reveal the degree to which attention distribution varies across people, content types, and even moments within a single exposure. For modeling purposes, eye tracking and dwell-time measures can be used to provide priors for possible asymmetric weights of the different modalities in the fusion model.

Attentional guidance is a specific form of sequential processing where one modality redirects processing toward another. Testing it requires detecting and encoding the presence of a relevant signal (e.g., language that says “see Figure 1”). Named entity recognition can identify explicit references to other modalities in linguistic content, and object detection tools can identify arrows, highlights, and other directional elements in visual content.

Then, once detected, the guided processing sequence should be encoded directionally before modeling downstream outcomes. Detections can be encoded as binary or categorical variables indicating both the presence of guidance and its direction (text-to-image vs. image-to-text) and incorporated as an interaction term in the regression between the binary guidance variable and the features of the guided modality. The guiding modality’s signal about what to attend to in the guided modality should be used to filter or re-weight the features extracted from the guided modality. If a caption says, “note the product’s color,” for example, color-related features should be up-weighted rather than treating all visual features equally.

Understanding What to Combine and When

So far, the paper focused on *how* to combine multiple modalities, but it's also important to consider *what* should be combined in the first place. Specifically, (1) whether multiple modalities *need* to be combined, and if so, (2) *which* modalities to combine, and (3) *which* features should be examined.

Is Multi- or Intermodal Analysis Always Necessary?

Given most marketing data is inherently multimodal, incorporating more modalities (and their interactions) often boosts understanding of the content and its consequences, and improves model fit and predictive accuracy (e.g., Luo et al. 2025). There are some situations, though, where intermodality is less likely to be relevant. If the context only involves one modality (e.g., emails that only contain language), for example, or if the analysis focuses on something mostly conveyed by one modality (e.g., a restaurant's service is conveyed mostly by review text, not images), unimodal analysis is likely sufficient. Similarly, if the goal is prediction or prescription, and one modality is already highly predictive, or generates sufficient improvement in decision outcome, it may be reasonable to focus on the most important modality.

When the context involves multiple modalities, though, it's important to consider which modalities to combine, and which features of those modalities to examine. While one may be tempted to use complex representations to try to completely represent the content of all potential modalities, for example, not all modalities, or features of them, are relevant for a given context or outcome, and may make things less interpretable. Consequently, it's important to move beyond convenience (i.e., what features can easily be measured) to a more conceptual understanding of what modalities and features are likely relevant in a given situation.

What Different Modalities Convey

Different modalities rely on different underlying systems, which have implications for what they effectively convey (see Table 5), and how relevant they may be in different contexts. Visual content, for example, can directly depict what something looks like (e.g., a picture of a tree), which enables it to effectively convey *visual and spatial information* (e.g., exactly what a product looks like or the layout of a shopping mall; Keller 2007; Kress 2005) and *proof* or *evidence* (e.g., a package is damaged). Similarly, by documenting things over time, multiple images in a row (i.e., video) can convey *movement and energy* (e.g., how excited a salesperson is).

Words, however, do not have to look like what they refer to (i.e., the word "shoe" doesn't actually depict a shoe; Saussure 1916), which allows language to represent things with no concrete form (e.g., fairness). This makes language particularly good at conveying *intangible information*, such as *abstract concepts* (e.g., sustainability or innovation), *internal thoughts* (e.g., an opinion about a brand), and motivations (e.g., "I bought this to get in shape"). In addition, because language follows grammatical rules that allow words to be combined to construct new meaning (e.g., saying "the car is red" creates the meaning that an object has a certain color; Halliday and Webster 2003; Larsen-Freeman 2011), it is particularly good at conveying *propositional content* such as *relational information* (e.g., "Product A is cheaper than product B"), *descriptive statements* (e.g., "these cost $129.99"), and *causal reasoning* (e.g., "it's the healthiest because we use the best ingredients"). Further, while vision and hearing focus on things that are currently present,[10] language can go *beyond the here and now* (Fitch 2010),

[10] While recording and media production have allowed vision and hearing to transcend the here and now (e.g., hearing recordings of the past or seeing potential futures in a movie), they are otherwise constrained by what can be represented and can only convey tangible, depictable content. In contrast, language supports abstractions and ideas that have no concrete form (e.g., counterfactuals) that are difficult to depict perceptually.

allowing it to convey the *past or future*, *temporal processes* (e.g., the steps to return a product), and *imagined possibilities*.

Hearing is a basic perceptual system that animals rely on to avoid predators and find prey, so it's not surprising that sounds (e.g., a fire alarm) can convey danger. Along these lines, audio can directly affect the nervous system (Salimpoor et al. 2011) and can be particularly good at conveying *emotion* or generating a quick emotional response. Further, given its immersive nature, audio can effectively convey *sensory richness* (e.g., the hiss of coffee machines) and the *atmosphere* that emerges from it (e.g., the café feels cozy).

Table 5: Examples of What Different Modalities Effectively Convey

Visual	• Visual and spatial information (e.g., what a product looks like, layout of a mall) • Proof or evidence (e.g., the package is damaged, the room looks clean) Specific to Video • Movement and energy (e.g., how excited a salesperson is)

Language	• Intangible information
	  ▪ Abstract concepts (e.g., this technology is innovative)
	  ▪ Internal thoughts (e.g., consumers' opinions about a brand)
	  ▪ Motivations (e.g., "I bought this to get in shape")
	• Propositional content
	  ▪ Relational information (e.g., "Product A is cheaper than product B")
	  ▪ Descriptive statements (e.g., "These cost $129.99")
	  ▪ Causal reasoning (e.g., "It's the healthiest because we use the best ingredients")
	• Beyond here and now
	  ▪ Temporal processes (e.g., steps to return a product)
	  ▪ Past, future, and imagined possibilities

Audio	• Emotion or quick emotional response (e.g., music or score of movie conveys excitement, or sets the tone) • Atmosphere (e.g., a store buzzing with activity)

The varying nature of different modalities also leads them to be *less* useful in conveying certain things. Because language uses abstract symbols to map onto meanings (e.g., the word "tree" to mean tree), for example, rather than depicting things directly (i.e., an image of a tree), it struggles to convey feelings of *scale* or *magnitude* (e.g., how big a space feels), and *movement* or

energy (e.g., saying "a performance feels alive" doesn't fully convey a concert's energy). Similarly, visual and audio content struggles to convey *intangible information* that has no visible or audible form, such as *abstract concepts* (e.g., free returns), *internal thoughts and motivations* (e.g., "I bought this gym membership to stay healthy"), and *propositional content* (e.g., "these eggs come from cage-free chickens").

These aspects suggest which modalities (and features) to examine. If the interest is reasoning (which tends to be conveyed by language), aesthetics (which tend to be conveyed visually), or atmosphere (which tends to be conveyed auditorily), then focusing on just the one relevant modality may be enough. If one is interested in consumer response to customer service calls, however, then it's likely important to consider both linguistic content (e.g., the information that was shared) and auditory content (e.g., the tone with which it was conveyed).

Understanding what each modality conveys also sets the stage for intermodality. When modalities differ in what they communicate, for example, one carrying propositional content, another conveying atmosphere or affect, their co-presence affects the form of intermodality (e.g., content-relatedness) they may generate. A customer service call where the agent says "I'm happy to help" in a flat tone, for example, is not just parallel linguistic and auditory events; it is precisely the *mismatch* between the two that shapes customer perception.

General Discussion

The marketplace is multimodal, and marketplace actors are constantly integrating linguistic, audio, and visual content. Further, digital markets have only increased the prevalence of such multimodal communication. But while a great deal of marketing research has analyzed language, and begun to analyze audio and visual content, there has been less attention to how

these different aspects combine.

To fill this gap, this paper brings together disparate findings and literatures to propose an integrative intermodality framework. Rather than treating modalities as independent streams, it describes how different modalities relate to, and interact with, one another and sheds light on *when* and *why* cross-modal interactions arise. Specifically, we outline three types of intermodality (i.e., content relatedness, interactive effects, and the nature of attention), as well as their underlying subtypes (e.g., content relatedness can involve content being aligned, complementary, contradictory, or unrelated).

We also lay out a unifying approach for modeling multimodal communications. This includes systematically aligning levels of representation, navigating tradeoffs between post-extraction interaction and joint modeling approaches, and specifying functional forms best suited to different types of intermodality. Taken together, these theoretical and methodological contributions provide concrete tools for moving from unimodal and multimodal analysis toward intermodality.

This work also has clear practical implications for a broad range of marketplace actors. Brands, call centers, and content creators can use the framework to better understand how to calibrate linguistic, auditory, and visual elements to maximize impact. The frameworks also provide guidance to researchers at every stage of the research process, including hypothesis generation, research design, data collection, feature extraction, and modeling. Rather than "looking under the streetlamp" for whatever type of data and features are most accessible, researchers can use the framework to identify which modalities are most theoretically relevant for a given question, which features to extract from each, how to align representation levels, and which functional forms best capture the type of cross-modal relationship they expect to find.

Directions for Future Research

The intermodality framework also highlights potential areas for future research. This includes (1) broadening the multimodal combinations studied, (2) identifying when different modalities have more impact, (3) broadening the information channels studied, and (4) extending intermodality across multiple pieces of content.

Broadening the multimodal combinations studied

Future research might broaden the multimodal combinations studied. Looking across existing multimodal work (Table 6) illustrates that (1) it is mostly restricted to two modalities, and (2) certain modality combinations have received more attention than others.

Table 6: Examples of Marketing Research Examining Multiple Modalities

	Linguistic and Visual Content	Linguistic and Audio Content	Visual and Audio Content	Linguistic, Audio, and Visual Content
Existing Work	• Villarroel Ordenes et al. (2019) • Lee (2021) • Troncoso and Luo (2023) • Cao, Li, and Zhang (2025) • Sikdar, Chakraborty, and Dogonadze (2025) • Zhang and Luo (2023) • Ceylan, Diehl, and Proserpio (2024) • Hartmann et al. (2021) • Li and Xie (2020) • Dew, Ansari, and Toubia (2022) • Farace et al. (2025) • Chung, Ding, and Kalra (2023) • Dang et al. (2026) • Yazdani, Chakravarty, and Inman (2025) • Zhou et al. (2021) • Kim, Jiang, and Thomadsen (2023) • Lin, Yao, and Chen (2021)	• Boughanmi and Ansari (2021) • Xu et al. (2025) • Cascio Rizzo and Berger (2023) • Balducci et al. (2026)	• Ding and Wang (2023)	• Chakraborty et al. (2025) • Tian, Dew, and Iyengar (2024) • Cascio Rizzo, Berger, and Zhou (2025) • Chang, Mukherjee, and Chattopadhyay (2023)
Marketplace Examples	Online reviews, social media posts, print and display ads, product descriptions	Customer calls, sales calls, radio, podcast ads, earning calls	YouTube videos with just music, ads without words	Livestream sales pitches, in person presentations, video ads, influencer videos

Most work has considered the intersection of language and visual content, for example, primarily in the context of social media (e.g., Li and Xie 2020) and online reviews (Ceylan, Diehl, and Proserpio 2024). Work leveraging video data has also studied the trimodal combination of linguistic, audio, and visual features (Bravin et al. 2025). Other modality pairs have received scant attention.

There could be several reasons for this uneven exploration. Some multimodalities may just be more prevalent. The combination of language and visual content appears frequently (e.g., in online reviews and social media posts), for example, but the combination of visual and audio is less common (e.g., videos of nature scenes).

Alternatively, even when modalities occur frequently in practice, some are more accessible, and researchers may be "looking under the streetlamp" (Berger et al. 2020; Du et al. 2021). Online reviews are more widely available than audio recordings, for example, which may partially explain the dearth of work on language and audio. Given how many marketplace interactions involve just language and audio features (e.g., customer service calls, sales calls, and word of mouth over the phone), though, this combination may deserve more attention.[11]

Development of analytic tools may also guide what has been studied. Tools for analyzing audio have lagged behind tools for text and image analysis, which may hinder work in this area.

Finally, as discussed, some modalities may be better than others at conveying particular types of information, which may drive attention to certain modality pairs. If a researcher is interested in *what* was communicated, for example, rather than *how*, nonlinguistic audio may be less informative than images or text.

[11] The difficulty of removing identifiable information from calls is one of the reasons for the limited availability of that form of data.

Regardless, being more aware of what has received more attention will hopefully encourage researchers to identify valuable areas to explore.

Which modalities have more impact?

Future work could also examine whether certain modalities are more likely to drive relevant outcomes (e.g., purchase or brand recall). One possibility is that certain modalities are just more impactful than others. Given linguistic content conveys semantic information, for example, such as facts and other propositional content, it may have a bigger impact on persuasion than audio or visual content. If so, marketers might want to spend more time optimizing the language of video advertisements than other dimensions.

Alternatively, which modalities play a larger role may depend on the context. Language's ability to provide facts and detailed reasons for action (i.e., "this is the best detergent because…") should be particularly useful for high involvement decisions (e.g., buying a car) or those that involve more central processing (i.e., deep, logical thinking). But when decisions are driven more by aesthetics (e.g., which flowers to buy), or require explicit proof (i.e., "we have the cleanest bathrooms"), visual information may be particularly impactful. Audio and visual content may also be particularly good at grabbing attention. A shout is more attention getting than its corresponding textual form (e.g., an exclamation mark), for example, and a fire alarm can immediately shift attention. Similarly, high contrast images (e.g., a bright product displayed against a dark background), or sudden visual onsets (e.g., pop-up ad) can attract attention automatically, even before viewers know what they are looking at. Visual content is also often more memorable than words (Paivio and Csapo 1973) and audio content seems particularly effective at evoking emotion.

Overall, then, which modality has a bigger impact may depend on the situation (and outcome) studied. This further highlights the importance of not only measuring and controlling for features of multiple modalities, but considering intermodality, and how modalities might interact to drive key outcomes.

Broadening the information channels studied

We focused on linguistic, auditory, and visual content, but future work might further divide these streams or explore additional information channels. Some work in communications and semiotics (e.g., New London Group 1996) separates visual content into aspects like gesturals (e.g., gestures and facial expressions) and other types of visual communication (e.g., charts). While there is certainly overlap between these aspects (e.g., both involve visual features like brightness or hue), there are also clear differences (e.g., certainty is more likely to be expressed by facial expressions than visual features of charts). Consequently, future work might examine both differences in what features to extract from these submodalities, and their potential effects.

Future work might also explore other channels of information such as smell, taste, and touch. While there is certainly experimental work on these modalities (Biswas and Szocs 2019; Hoegg and Alba 2007; Krishna, Elder, and Caldara 2010; Luangrath et al. 2022), we are not aware of any papers that have analyzed unstructured data of this type. That does not mean that these modalities aren't important. In fact, one of the first things consumers may notice when walking into a restaurant is the smell. The sparse research on these modalities may be driven by lack of data (at least at the moment), lack of appropriate methods,[12] or a tendency to study the

[12] Other fields have begun to develop models to predict sensory perceptions from physical features. Work in chemistry (e.g., Dagan-Wiener et al. 2017; Lee et al. 2023), for example, uses chemical structure to predict odor and taste, and work in computer science (Awan and Jeon 2025; Hassan, Joolee, and Jeon 2023) uses surface image

modalities that are easiest to study. Regardless, future work could examine how olfactory, tactile, and gustatory features impact consumer behavior.

Extending Intermodality Across Multiple Pieces of Content

While this paper focused on extracting multiple modalities from a single piece of content (e.g., linguistic and visual features from an online review), the frameworks and tools we propose should also be valuable in situations where people consume multiple pieces of related content. Greater similarity between a digital ad's visual content, for example, and the linguistic content of the landing page it links to (Kim and Kalyanam 2025), might encourage purchase. Similarly, the similarity of a thumbnail image to the language used in the video it leads to may shape consumer engagement (Yang and Netzer 2026). As long as the multiple pieces of content are related, and their relationship shapes some downstream aspect of consumer behavior, multimodal analysis can help answer relevant marketing questions. Indeed, such examples can be seen as forms of sequential processing discussed earlier.

Conclusion

In conclusion, given the ubiquity of multimodal content, we hope that this work will encourage more researchers to not only go beyond unimodality to multimodality, but to think about intermodality more broadly.

---

and/or tactile sensor data to predict perceived textures (e.g., rough or smooth). These efforts are similar to how consumer research uses text, audio, and images to infer psychological states and predict perceptions.

References

Al-Zoghby, Aya M., Esraa Mohamed Al-Awadly, Ahmed Ismail Ebada, and Wael A. Awad (2025), "Overview of Multimodal Machine Learning," *ACM Transactions on Asian and Low-Resource Language Information Processing*, 24 (1), 1–20.

Alais, David and David Burr (2004), "The Ventriloquist Effect Results From Near-Optimal Bimodal Integration," *Current Biology*, 14 (3), 257–62.

Arora, Neeraj, Ishita Chakraborty, and Yohei Nishimura (2025), "AI–Human Hybrids for Marketing Research: Leveraging Large Language Models (LLMs) as Collaborators," *Journal of Marketing*, 89 (2), 43–70.

Athey, Susan (2018), "The Impact of Machine Learning on Economics," in *The Economics of Artificial Intelligence: An Agenda*, Ajay K. Agrawal, Joshua Gans, and Avi Goldfarb, eds. University of Chicago Press, 507–47.

Awan, Mudassir Ibrahim and Seokhee Jeon (2025), "Estimating Perceptual Attributes of Haptic Textures Using Visuo-Tactile Data," IEEE Access, https://doi.org/10.1109/ACCESS.2025.3581685.

Balducci, Bitty, Bin Pang, Lingshu Hu, Can Li, Wenbo Wang, Yi Shang, Detelina Marinova, and Matt Gordon (2026), "Leveraging Audio Data: A Guide to Understanding Customer-Firm Conversations," *Marketing Letters*, 37 (1), 10.

Baltrusaitis, Tadas, Chaitanya Ahuja, and Louis-Philippe Morency (2018), "Multimodal Machine Learning: A Survey and Taxonomy," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41 (2), 423–43.

Bengio, Yoshua, Aaron Courville, and Pascal Vincent (2013), "Representation Learning: A Review and New Perspectives," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35 (8), 1798–828.

Berger, Jonah, Ashlee Humphreys, Stephan Ludwig, Wendy W. Moe, Oded Netzer, and David A. Schweidel (2020), "Uniting the Tribes: Using Text for Marketing Insight," *Journal of Marketing*, 84 (1), 1–25.

Berger, Jonah and Katherine L. Milkman (2012), "What Makes Online Content Viral?," *Journal of Marketing Research*, 49 (2), 192–205.

Berger, Jonah, Wendy W. Moe, and David A. Schweidel (2023), "What Holds Attention? Linguistic Drivers of Engagement," *Journal of Marketing*, 87 (5), 793–809.

Berger, Jonah and Grant Packard (2018), "Are Atypical Things More Popular?," *Psychological Science*, 29 (7), 1178–84.

Berlyne, Daniel E. (1960), *Conflict, Arousal, and Curiosity*, McGraw-Hill.

Biswas, Dipayan and Courtney Szocs (2019), "The Smell of Healthy Choices: Cross-Modal Sensory Compensation Effects of Ambient Scent on Food Purchases," *Journal of Marketing Research*, 56 (1), 123–41.

Blanchard, Simon J., Nofar Duani, Aaron M. Garvey, Oded Netzer, and Travis Tae Oh (2025), "New Tools, New Rules: A Practical Guide to Effective and Responsible Generative AI Use for Surveys and Experiments in Research," *Journal of Marketing*, 89 (6), 119–39.

Boughanmi, Khaled and Asim Ansari (2021), "Dynamics of Musical Success: A Machine Learning Approach for Multimedia Data Fusion," *Journal of Marketing Research*, 58 (6), 1034–57.

Boyd, Ryan L., Ashwini Ashokkumar, Sarah Seraj, and James W. Pennebaker (2022), "The Development and Psychometric Properties of LIWC-22," technical report, University of Texas at Austin.

Bravin, Marc, Melanie Clegg, Reto Hofstetter, Marc Pouly, and Jonah A. Berger (2025), "How to Follow Social Media Trends? An Empirical Investigation Using TikTok Short Video Data," SSRN (February 24), https://doi.org/10.2139/ssrn.5117564.

Bucher, Hansjürgen and Peter Schumacher (2006), "The Relevance of Attention for Selecting News Content. An Eye-Tracking Study on Attention Patterns in the Reception of Print and Online Media," *Communications*, 31 (3), 347.

Cao, Jingcun, Xiaolin Li, and Lingling Zhang (2025), "Is Relevancy Everything? A Deep-Learning Approach to Understand the Effect of Image-Text Congruence," *Management Science*, 71 (12), 10579–602.

Cascio Rizzo, Giovanni Luca and Ignazio Ziano (2025), "Response Speed Influences Consumer Evaluations," SSRN (June 13), https://doi.org/10.2139/ssrn.5286938.

Cascio Rizzo, Giovanni Luca, Jonah Berger, and Mi Zhou (2025), "Talking with Your Hands: How Hand Gestures Influence Communication," *Journal of Marketing Research*, published online September 25, https://doi.org/10.1177/00222437251385922.

Cascio Rizzo, Giovanni Luca, Jonah Berger, and Francisco Villarroel Ordenes (2023), "What Drives Virtual Influencer's Impact?," arXiv (January 24), https://doi.org/10.48550/arXiv.2301.09874.

Cascio Rizzo, Giovanni Luca and Jonah A. Berger (2023), "The Power of Speaking Slower," SSRN (October 19), https://doi.org/10.2139/ssrn.4580994.

Ceylan, Gizem, Kristin Diehl, and Davide Proserpio (2024), "Words Meet Photos: When and Why Photos Increase Review Helpfulness," *Journal of Marketing Research*, 61 (1), 5–26.

Chakraborty, Ishita, Khai Chiong, Howard Dover, and K. Sudhir (2025), "Can AI and AI-Hybrids Detect Persuasion Skills? Salesforce Hiring With Conversational Video Interviews," *Marketing Science*, 44 (1), 30–53.

Chang, Hannah H., Anirban Mukherjee, and Amitava Chattopadhyay (2023), "More Voices Persuade: The Attentional Benefits of Voice Numerosity," *Journal of Marketing Research*, 60 (4), 687–706.

Chung, Jaeyeon (Jae), Yu Ding, and Ajay Kalra (2023), "I Really Know You: How Influencers Can Increase Audience Engagement by Referencing Their Close Social Ties," *Journal of Consumer Research*, 50 (4), 683–703.

Cummings, Alycia, Rita Čeponienė, Frederic Dick, Ayse Pinar Saygin, and Jeanne Townsend (2008), "A Developmental ERP Study of Verbal and Non-Verbal Semantic Processing," *Brain Research*, 1208, 137–49.

Dagan-Wiener, Ayana, Ido Nissim, Natalie Ben Abu, Gigliola Borgonovo, Angela Bassoli, and Masha Y. Niv (2017), "Bitter or Not? Bitterpredict, a Tool for Predicting Taste From Chemical Structure," *Scientific Reports*, 7 (1).

Dang, Ivy Chu, Canice M.C. Kwan, Jayson S. Jia, and Yang Shi (2026), "When Words Meet Visuals: How Content Composition Drives Social Media Engagement for Marketer-Generated Content," *Journal of Marketing Research*, 63 (1), 167–90.

Dew, Ryan, Asim Ansari, and Olivier Toubia (2022), "Letting Logos Speak: Leveraging Multiview Representation Learning for Data-Driven Branding and Logo Design," *Marketing Science*, 41 (2), 401–25.

Ding, MengQi Annie and Xin Shane Wang (2023), "The Effect of Image-Audio Emotional Similarity on NFT Product Sales," SSRN (June 28), https://ssrn.com/abstract=4492032.
Du, Rex Yuxing, Oded Netzer, David A. Schweidel, and Debanjan Mitra (2021), "Capturing Marketing Information to Fuel Growth," *Journal of Marketing*, 85 (1), 163–83.
Dzyabura, Daria and Renana Peres (2021), "Visual Elicitation of Brand Perception," *Journal of Marketing*, 85 (4), 44–66.
Efthymiou, Fotis, Christian Hildebrand, Emanuel de Bellis, and William H. Hampton (2023), "The Power of AI-Generated Voices: How Digital Vocal Tract Length Shapes Product Congruency and Ad Performance," *Journal of Interactive Marketing*, 59 (2), 117–34.
Exner, Yannick, Jochen Hartmann, Oded Netzer, Shunyuan Zhang, and Ziqian Ding (2025), "AI in Disguise: Quasi-Experimental Analysis of a Large-Scale Deployment of AI-Generated Display Ads," SSRN (March 7), https://doi.org/10.2139/ssrn.5096969.
Farace, Stefania, Francisco Villarroel Ordenes, Dennis Herhausen, Dhruv Grewal, and Ko de Ruyter (2025), "Standing Out While Fitting In: Visual Design of Text Overlays in Social Media Communication," *Journal of Marketing*, 90 (1), 132–51.
Feng, Xiaohang (Flora), Charis X Li, and Shunyuan Zhang (2025), "Visual Uniqueness in Peer-To-Peer Marketplaces: Machine Learning Model Development, Validation, and Application," *Journal of Consumer Research*, 52 (4), 800–25.
Feng, Xiaohang, Shunyuan Zhang, Xiao Liu, Kannan Srinivasan, and Cait Lamberton (2025), "An AI Method to Score Celebrity Visual Potential," *Journal of Marketing Research*, 62 (5), 757–75.
Fitch, Tecumseh (2010), The Evolution of Language, Cambridge University Press.
Fleming, Neil D. and Colleen Mills (1992), "Not Another Inventory, Rather a Catalyst for Reflection," *To Improve the Academy*, 11 (1), 137–55.
Fong, Hortense, Vineet Kumar, and K. Sudhir (2025), "A Theory-Based Explainable Deep Learning Architecture for Music Emotion," *Marketing Science*, 44 (1), 196–219.
Forceville, Charles (1996), *Pictorial Metaphor in Advertising*, London: Routledge.
Gillis, Randall L. and Elizabeth S. Nilsen (2017), "Consistency Between Verbal and Non-Verbal Affective Cues: A Clue to Speaker Credibility," *Cognition and Emotion*, 31 (4), 645–56.
Grewal, Rajdeep, Sachin Gupta, and Rebecca Hamilton (2021), "Marketing Insights From Multimedia Data: Text, Image, Audio, and Video," *Journal of Marketing Research*, 58 (6), 1025–33.
Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart (2021), "Machine Learning for Social Science: An Agnostic Approach," *Annual Review of Political Science*, 24 (1), 395–419.
Hagoort, Peter (2013), "MUC (Memory, Unification, Control) and Beyond," *Frontiers in Psychology*, 4, 416.
Halliday, M. A. K. and Jonathan J. Webster (2003), *On Language and Linguistics*, London: Bloomsbury Publishing.
Hartmann, Jochen, Yannick Exner, and Samuel Domdey (2025), "The Power of Generative Marketing: Can Generative AI Create Superhuman Visual Marketing Content?" *International Journal of Research in Marketing*, 42 (1), 13–31.
Hartmann, Jochen, Mark Heitmann, Christina Schamp, and Oded Netzer (2021), "The Power of Brand Selfies," *Journal of Marketing Research*, 58 (6), 1159–77.

Hartmann, Jochen, Mark Heitmann, Christian Siebert, and Christina Schamp (2023), "More Than a Feeling: Accuracy and Application of Sentiment Analysis," *International Journal of Research in Marketing*, 40 (1), 75–87.
Hartmann, Jochen and Oded Netzer (2023), "Natural Language Processing in Marketing," in *Artificial Intelligence in Marketing*, K. Sudhir and Olivier Toubia, eds.
Hassan, Waseem, Joolekha Bibi Joolee, and Seokhee Jeon (2023), "Establishing Haptic Texture Attribute Space and Predicting Haptic Attributes From Image Features Using 1D-Cnn," *Scientific Reports*, 13 (1).
Heckler, Susan E. and Terry L. Childers (1992), "The Role of Expectancy and Relevancy in Memory for Verbal and Visual Information: What Is Incongruency?," *Journal of Consumer Research*, 18 (4), 475.
Hildebrand, Christian, Fotis Efthymiou, Francesc Busquet, William H. Hampton, Donna L. Hoffman, and Thomas P. Novak (2020), "Voice Analytics in Business Research: Conceptual Foundations, Acoustic Feature Extraction, and Applications," *Journal of Business Research*, 121, 364–74.
Hochreiter, Sepp and Jürgen Schmidhuber (1997), "Long Short-Term Memory," *Neural Computation*, 9 (8), 1735–80.
Hockett, Charles F. (1960), "The Origin of Speech," Scientific American, 203 (3), 88–96.
Hoegg, JoAndrea and Joseph W. Alba (2007), "Taste Perception: More Than Meets the Tongue," *Journal of Consumer Research*, 33 (4), 490–98.
Hong, Jiyeon and Paul R. Hoban (2022), "Writing More Compelling Creative Appeals: A Deep Learning-Based Approach," *Marketing Science*, 41 (5), 941–65.
Huang, Mengyao and Zsolt Katona (2025), "Consumer Evaluation and Visual Attribution of Realistic AI-Generated Advertisements," SSRN (July 30), https://doi.org/10.2139/ssrn.5366745.
Humphreys, Ashlee and Rebecca Jen-Hui Wang (2017), "Automated Text Analysis for Consumer Research," *Journal of Consumer Research*, 44 (6), 1274–306.
Jackendoff, Ray (2002), *Foundations of Language: Brain, Meaning, Grammar, Evolution*, Oxford: Oxford University Press.
Keller, Daniel (2007), "Thinking Rhetorically," in *Multimodal Composition: Resources for Teachers*, Cynthia L. Selfe, ed. Hampton Press, 49–63.
Kim, Donggwan, Zhenling Jiang, and Raphael Thomadsen (2023), "TV Advertising Effectiveness with Racial Minority Representation: Evidence from the Mortgage Market," SSRN (July 27), https://doi.org/10.2139/ssrn.4521178.
Kim, Yewon and Kirthi Kalyanam (2025), "Are Display Ad Content and Landing Page Ad Content Complements or Substitutes? A Field Experiment," SSRN (October 19), https://doi.org/10.2139/ssrn.5620810.
Kress, Gunther (2005), "Gains and Losses: New Forms of Texts, Knowledge, and Learning," *Computers and Composition*, 22 (1), 5–22.
Kress, Gunther and Theo Van Leeuwen (2001), *Multimodal Discourse: The Modes and Media of Contemporary Communication*, London: Arnold Publishers.
Krishna, Aradhna, Ryan S. Elder, and Cindy Caldara (2010), "Feminine to Smell but Masculine to Touch? Multisensory Congruence and Its Effect on the Aesthetic Experience," *Journal of Consumer Psychology*, 20 (4), 410–18.

Kutas, Marta and Kara D. Federmeier (2011), "Thirty Years and Counting: Finding Meaning in the N400 Component of the Event-Related Brain Potential (ERP)," *Annual Review of Psychology*, 62 (1), 621–47.

Larsen-Freeman, Diane (2011), "Key Concepts in Language Learning and Language Education," in *The Routledge Handbook of Applied Linguistics*, James Simpson, ed. Routledge, 155–70.

LeDoux, Joseph E. (1996), *The Emotional Brain: The Mysterious Underpinnings of Emotional Life*, Simon and Schuster.

Lee, Brian K., Emily J. Mayhew, Benjamin Sanchez-Lengeling, Jennifer N. Wei, Wesley W. Qian, Kelsie A. Little, Matthew Andres, Britney B. Nguyen, Theresa Moloy, Jacob Yasonik, Jane K. Parker, Richard C. Gerkin, Joel D. Mainland, and Alexander B. Wiltschko (2023), "A Principal Odor Map Unifies Diverse Tasks in Olfactory Perception," *Science*, 381 (6661), 999–1006.

Lee, Jeffrey K. (2021), "Emotional Expressions and Brand Status," *Journal of Marketing Research*, 58 (6), 1178–96.

Li, Hongshuang (Alice) and Liye Ma (2020), "Charting the Path to Purchase Using Topic Models," *Journal of Marketing Research*, 57 (6), 1019–36.

Li, Yiyi and Ying Xie (2020), "Is a Picture Worth a Thousand Words? An Empirical Study of Image Content and Social Media Engagement," *Journal of Marketing Research*, 57 (1), 1–19.

Lin, Yan, Dai Yao, and Xingyu Chen (2021), "Happiness Begets Money: Emotion and Engagement in Live Streaming," *Journal of Marketing Research*, 58 (3), 417–38.

Liu, Xuan, Savannah Wei Shi, Thales Teixeira, and Michel Wedel (2018), "Video Content Marketing: The Making of Clips," *Journal of Marketing*, 82 (4), 86–101.

Luangrath, Andrea Webb, Joann Peck, William Hedgcock, and Yixiang Xu (2022), "Observing Product Touch: The Vicarious Haptic Effect in Digital Marketing and Virtual Reality," *Journal of Marketing Research*, 59 (2), 306–26.

Matz, Sandra, Cristina Segalin, David Stillwell, Sandrine R. Müller, and Maarten W. Bos (2019), "Predicting the Personal Appeal of Marketing Images Using Computational Methods," *Journal of Consumer Psychology*, 29 (3), 370–90.

Mayer, Richard E. (2001), *Multimedia Learning*, Cambridge: Cambridge University Press.

Mayer, Richard E. and Richard B. Anderson (1992), "The Instructive Animation: Helping Students Build Connections Between Words and Pictures in Multimedia Learning," *Journal of Educational Psychology*, 84 (4), 444.

McGurk, Harry and John MacDonald (1976), "Hearing Lips and Seeing Voices," *Nature*, 264 (5588), 746–48.

Meyers-Levy, Joan and Alice M. Tybout (1989), "Schema Congruity as a Basis for Product Evaluation," *Journal of Consumer Research*, 16 (1), 39.

Morency, Louis-Philippe and Tadas Baltrušaitis (2017), "Multimodal Machine Learning: Integrating Language, Vision and Speech," *Proceedings of ACL 2017*, Tutorial Abstracts, 3–5.

Moreno, Roxana and Richard E. Mayer (1999), "Cognitive Principles of Multimedia Learning: The Role of Modality and Contiguity," *Journal of Educational Psychology*, 91 (2), 358.

Nanne, Annemarie J., Marjolijn L. Antheunis, Chris G. Van Der Lee, Eric O. Postma, Sander Wubben, and Guda Van Noort (2020), "The Use of Computer Vision to Analyze Brand-

Related User Generated Image Content," *Journal of Interactive Marketing*, 50 (1), 156–67.
Netzer, Oded, Alain Lemaire, and Michal Herzenstein (2019), "When Words Sweat: Identifying Signals for Loan Default in the Text of Loan Applications," *Journal of Marketing Research*, 56 (6), 960–80.
Netzer, Oded, Ronen Feldman, Jacob Goldenberg, and Moshe Fresko (2012), "Mine Your Own Business: Market-Structure Surveillance Through Text Mining," *Marketing Science*, 31 (3), 521–43.
New London Group (1996), "A Pedagogy of Multiliteracies: Designing Social Futures," *Harvard Educational Review*, 66 (1), 60–93.
Öhman, Arne (2008), "Fear and Anxiety: Overlaps and Dissociations," in *Handbook of Emotions*, 3rd ed., Michael Lewis, Jeannette M. Haviland-Jones, and Lisa Feldman Barrett, eds. The Guilford Press, 709–28.
Packard, Grant and Jonah Berger (2020), "Thinking of You: How Second-Person Pronouns Shape Cultural Success," *Psychological Science*, 31 (4), 397–407.
Packard, Grant and Jonah Berger (2024), "The Emergence and Evolution of Consumer Language Research," *Journal of Consumer Research*, 51 (1), 42–51.
Paivio, Allan (1971), *Imagery and Verbal Processes*, Holt, Rinehart & Winston.
Paivio, Allan and Kalman Csapo (1973), "Picture Superiority in Free Recall: Imagery or Dual Coding?," *Cognitive Psychology*, 5 (2), 176–206.
Partan, Sarah and Peter Marler (1999), "Communication Goes Multimodal," *Science*, 283 (5406), 1272–73.
Pieters, Rik and Michel Wedel (2004), "Attention Capture and Transfer in Advertising: Brand, Pictorial, and Text-Size Effects," *Journal of Marketing*, 68 (2), 36–50.
Potter, Mary C. (1976), "Short-Term Conceptual Memory for Pictures," *Journal of Experimental Psychology: Human Learning and Memory*, 2 (5), 509–22.
Rathje, Steve, Dan-Mircea Mirea, Ilia Sucholutsky, Raja Marjieh, Claire E. Robertson, and Jay J. Van Bavel (2024), "GPT Is an Effective Tool for Multilingual Psychological Text Analysis," *Proceedings of the National Academy of Sciences*, 121 (34), e2308950121.
Rayner, Keith, Caren M. Rotello, Andrew J. Stewart, Jessica Keir, and Susan A. Duffy (2001), "Integrating Text and Pictorial Information: Eye Movements When Looking at Print Advertisements," *Journal of Experimental Psychology: Applied*, 7 (3), 219–26.
Rocklage, Matthew D., Sharlene He, Derek D. Rucker, and Loran F. Nordgren (2023), "Beyond Sentiment: The Value and Measurement of Consumer Certainty in Language," *Journal of Marketing Research*, 60 (5), 870–88.
Rocklage, Matthew D., Derek D. Rucker, and Loran F. Nordgren (2018), "The Evaluative Lexicon 2.0: The Measurement of Emotionality, Extremity, and Valence in Language," *Behavior Research Methods*, 50 (4), 1327–44.
Ruiz, Francisco J.R., Susan Athey, and David M. Blei (2020), "Shopper," *The Annals of Applied Statistics*, 14 (1), 1–27.
Salimpoor, Valorie N., Mitchel Benovoy, Kevin Larcher, Alain Dagher, and Robert J. Zatorre (2011), "Anatomically Distinct Dopamine Release During Anticipation and Experience of Peak Emotion to Music," *Nature Neuroscience*, 14 (2), 257–62.
Saussure, Ferdinand de (1916), "Nature of the Linguistic Sign," in *Course in General Linguistics*, 65–70.

Sikdar, Sharmistha, Ishita Chakraborty, and Nika Dogonadze (2025), "Neither a Picasso nor a Leonardo da Vinci: An Examination of Novice Artwork Pricing with Multimodal Data," *Journal of Marketing*, published online December 5, https://doi.org/10.1177/00222429251408346.
Simonov, Andrey, Tommaso Valletti, and Andre Veiga (2025), "Attention Spillovers from News to Ads: Evidence from an Eye-Tracking Experiment," *Journal of Marketing Research*, 62 (2), 294–315.
Stein, Barry E. and M. Alex Meredith (1993), *The Merging of the Senses*, Cambridge, MA: MIT Press.
Thorpe, Simon, Denis Fize, and Catherine Marlot (1996), "Speed of Processing in the Human Visual System," *Nature*, 381 (6582), 520–22.
Tian, Zijun, Ryan Dew, and Raghuram Iyengar (2024), "Mega or Micro? Influencer Selection Using Follower Elasticity," *Journal of Marketing Research*, 61 (3), 472–95.
Timoshenko, Artem and John R. Hauser (2019), "Identifying Customer Needs From User-Generated Content," *Marketing Science*, 38 (1), 1–20.
Tirunillai, Seshadri and Gerard J. Tellis (2014), "Mining Marketing Meaning From Online Chatter: Strategic Brand Analysis of Big Data Using Latent Dirichlet Allocation," *Journal of Marketing Research*, 51 (4), 463–79.
Toubia, Olivier, Jonah Berger, and Jehoshua Eliashberg (2021), "How Quantifying the Shape of Stories Predicts Their Success," *Proceedings of the National Academy of Sciences*, 118 (26), e2011695118.
Troncoso, Isamar and Lan Luo (2023), "Look the Part? The Role of Profile Pictures in Online Labor Markets," *Marketing Science*, 42 (6), 1080–100.
Villarroel Ordenes, Francisco, Dhruv Grewal, Stephan Ludwig, Ko De Ruyter, Dominik Mahr, and Martin Wetzels (2019), "Cutting Through Content Clutter: How Speech and Image Acts Drive Consumer Sharing of Social Media Brand Messages," *Journal of Consumer Research*, 45 (5), 988–1012.
Villarroel Ordenes, Francisco and Shunyuan Zhang (2019), "From Words to Pixels: Text and Image Mining Methods for Service Research," *Journal of Service Management*, 30 (5), 593–620.
Wang, Xin (Shane), Neil Bendle, and Yinjie Pan (2024), "Beyond Text: Marketing Strategy in a World Turned Upside Down," *Journal of the Academy of Marketing Science*, 52, 939–54.
Wang, Xin (Shane), Shijie Lu, X I Li, Mansur Khamitov, and Neil Bendle (2021), "Audio Mining: The Role of Vocal Tone in Persuasion," *Journal of Consumer Research*, 48 (2), 189–211.
Xu, Haifeng, Yi Ding, Yu Ding, Qi Zhang, and Cheng Zhang (2025), "Whispers in Your Mind: The Role of Voice Features in Customer Acquisition and Retention," *Journal of Marketing*, 90 (1), 29–50.
Yang, Jasmine and Oded Netzer (2026), "What Makes for a Good Thumbnail? Video Content Summarization Into a Single Image," *Working paper*.
Yang, Joonhyuk, Yingkang Xie, Lakshman Krishnamurthi, and Purushottam Papatla (2022), "High-Energy Ad Content: A Large-Scale Investigation of TV Commercials," *Journal of Marketing Research*, 59 (4), 840–59.
Yazdani, Elham, Anindita Chakravarty, and Jeff Inman (2025), "(Mis)Alignment Between Facial and Textual Emotions and Its Effects on Donors Engagement Behavior in Online

Crowdsourcing Platforms," *Journal of the Academy of Marketing Science*, 53 (4), 968–88.

Zhang, Mengxia and Lan Luo (2023), "Can Consumer-Posted Photos Serve as a Leading Indicator of Restaurant Survival? Evidence From Yelp," *Management Science*, 69 (1), 25–50.

Zhang, Shunyuan, Elizabeth M.S. Friedman, Kannan Srinivasan, Ravi Dhar, and Xupin Zhang (2025), "Serving with a Smile on Airbnb: Analyzing the Economic Returns and Behavioral Underpinnings of the Host's Smile," *Journal of Consumer Research*, 51 (6), 1073–97.

Zhou, Mi, George H. Chen, Pedro Ferreira, and Michael D. Smith (2021), "Consumer Behavior in the Online Classroom: Using Video Analytics and Machine Learning to Understand the Consumption of Video Courseware," *Journal of Marketing Research*, 58 (6), 1079–100.

Zhu, Yuting, Xinyu Cao, Yuzhuo Su, and Yongbin Ma (2025), "Measuring Information Richness in Product Images: Implications for Online Sales," working paper, arXiv:2508.04541, https://arxiv.org/abs/2508.04541.

Appendix A: Unimodal Research in Marketing

Over the past 15 years, a burgeoning stream of literature has used automated textual analysis to study how language reflects and shapes consumer behavior (see Berger et al. 2020; Hartmann and Netzer 2023; Humphreys and Wang 2017; Packard and Berger 2024 for reviews). Given the rise of user generated content (e.g., consumer reviews and social media posts), and the fact that it is often readily accessible for analysis (Matz and Netzer 2017), existing work has focused on this type of data, extracting aspects like sentiment, certainty, and product features (e.g., Lee and Bradlow 2011; Netzer et al. 2012; Rocklage, Berger, and Boghrati 2025; Rocklage and Fazio 2020, see Table W1 for examples). Other work has examined the language of advertisements, news articles, loan applications, Airbnb host motivations, customer service calls, and movie scripts, all with the goal of better understanding consumer behavior and marketplace outcomes (e.g., Berger and Milkman 2012; Berger and Packard 2018; Chung et al. 2022; Lee, Hosanagar, and Nair 2018; Netzer, Lemaire, and Herzenstein 2019; Packard and Berger 2021; Toubia, Berger, and Eliashberg 2021). Indeed, the past decade had seen an over 300% increase in consumer language research (Packard and Berger 2024).

Many marketplace interactions (e.g., customer service calls and television advertisements) also contain audio features (e.g., pitch, tone, and melody) and advances in automated audio analytics have spurred research in this area (see Hildebrand et al. 2020 for reviews). This includes auditory properties such as stress, focus, brightness, loudness (Wang et al. 2021; Xu et al. 2025, see Table W1), as well as the emotional responses audio evokes (Fong, Kumar, and Sudhir 2025). Researchers have investigated audio across a wide range of contexts, including advertising (e.g., video ads and television commercials; Chang, Mukherjee, and Chattopadhyay 2023; Yang et al. 2022), customer-firm interactions (e.g., customer-service and sales calls; Balducci et al. 2026; Cascio Rizzo and Berger 2023; Xu et al. 2025), crowdfunding communication (e.g., pitch videos; Wang et al. 2021), and music (Fong, Kumar, and Sudhir 2025).

Bolstered by recent advances in computer vision, the popularity of image-based social media platforms (e.g., Instagram), and wider availability of video data, research has also started to study how visual features shape consumer perceptions and behavior (see Dzyabura and Peres 2021 for a review). Commonly studied image features include low-level visual features like color variation and compositional aspects, or facial presence and expression, number of people, and brand appearance (Chung, Ding, and Kalra 2023; Hartmann et al. 2021; Lee 2021; Li and Xie 2020; Matz et al. 2019, see Table W1). These aspects have been analyzed across a range of image types including user-generated content (Feng, Li, and Zhang 2025; Hartmann et al. 2021; Zhang and Luo 2023), brand-generated content (e.g., visual logos and social media posts, Dew, Ansari, and Toubia 2022; Hartmann et al. 2021; Lee 2021; Li and Xie 2020), and facial images (Feng et al. 2025).

Research is also beginning to examine video data, which contains a sequence of images, and often includes language and audio features (see Schramla 2025 and Schwenzow et al. 2021 for reviews). This work has studied aspects of motion (e.g., magnitude), color, and nonverbal cues such as hand movements and other body language (Cascio Rizzo, Berger, and Zhou 2025; Chakraborty et al. 2025; Zhou et al. 2021). Existing research has analyzed video across a variety of contexts, including educational videos (e.g., TED talks and online courses; Cascio Rizzo, Berger, and Zhou 2025; Zhou et al. 2021), livestream retails (Bharadwaj et al. 2022; Yang, Zhang, and Zhang 2025), and interview settings (Chakraborty et al. 2025).

Table W1: Example Features Studied in Different Modalities

	Linguistic Content	Audio Content	Visual Content	
Common data sources	Online reviews, Social media posts, customer service calls, sales pitches, livestreams, search ads	Customer service calls, sales pitches, livestreams, influencer marketing	Images Online reviews, Digital ads, social media posts	Video Livestreams, Digital ads, social media posts influencer marketing videos
Features Studied	• Valence (Berger and Milkman 2012; Melumad, Inman, and Pham 2019; Rocklage and Fazio 2015) • Concreteness (Humphreys, Isaac, and Wang 2021; Packard and Berger 2021; Umashankar, Kim, and Reutterer 2023) • Arousal (Cascio Rizzo, Berger, and Rocklage 2024; Kanuri, Chen, and Sridhar 2018; Yin, Bond, and Zhang 2017) • Warmth (Marinova, Singh, and Singh 2018; Packard, Li, and Berger 2023) • Certainty (Cascio Rizzo, Berger, and Rocklage 2024; Chakraborty et al. 2025; Rocklage, Berger, and Boghrati 2025) • Familiarity (Berger, Moe, and Schweidel 2023; Cascio Rizzo, Berger, and Villarroel Ordenes 2023; Mosley, Schweidel, and Zhang 2024) • Motivations (Chung et al. 2022)	• Loudness (Chang, Mukherjee, and Chattopadhyay 2023; Xu et al. 2025) • Pitch (Balducci et al. 2026; Fong, Kumar, and Sudhir 2025) • Tone (Cascio Rizzo and Berger 2023; Wang et al. 2021) • Brightness (Chakraborty et al. 2025; Wang et al. 2021; Xu et al. 2025) • Harmonics (Fong, Kumar, and Sudhir 2025) • Competence (Wang et al. 2021)	Images • Color (Dew, Ansari, and Toubia 2022; Matz et al. 2019) • Specific emotion (Li and Xie 2020; Troncoso and Luo 2023) • Semantic meaning (Cao, Li, and Zhang 2025; Ceylan, Diehl, and Proserpio 2024; Hartmann et al. 2021) • Arousal (Matz et al. 2019) • Emotionality (Lee 2021)	Video • Motion features (Zhou et al. 2021) • Scene cut (Liu et al. 2018; Zhou et al. 2021) • Body language (Cascio Rizzo, Berger, and Zhou 2025; Chakraborty et al. 2025) • Specific emotion (Bharadwaj et al. 2022; McDuff and Berger 2020; Yang, Zhang, and Zhang 2025; Zhou et al. 2021)

References

Bharadwaj, Neeraj, Michel Ballings, Prasad A. Naik, Miller Moore, and Mustafa Murat Arat (2022), "A New Livestream Retail Analytics Framework to Assess the Sales Impact of Emotional Displays," *Journal of Marketing*, 86 (1), 27–47.

Cascio Rizzo, Giovanni Luca, Jonah A. Berger, and Matthew Rocklage (2024), "How Speaking and Writing Shape Certainty," The Wharton School Research Paper, SSRN (June 4), https://doi.org/10.2139/ssrn.4854650.

Chung, Jaeyeon, Gita Venkataramani Johar, Yanyan Li, Oded Netzer, and Matthew Pearson (2022), "Mining Consumer Minds: Downstream Consequences of Host Motivations for Home-Sharing Platforms," *Journal of Consumer Research*, 48 (5), 817–38.

Humphreys, Ashlee, Mathew S. Isaac, and Rebecca Jen-Hui Wang (2021), "Construal Matching in Online Search: Applying Text Analysis to Illuminate the Consumer Decision Journey," *Journal of Marketing Research*, 58 (6), 1101–19.

Kanuri, Vamsi K., Yixing Chen, and Shrihari Sridhar (2018), "Scheduling Content on Social Media: Theory, Evidence, and Application," *Journal of Marketing*, 82 (6), 89–108.

Lee, Dokyun, Kartik Hosanagar, and Harikesh S. Nair (2018), "Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook," *Management Science*, 64 (11), 5105–31.

Lee, Thomas and Eric Bradlow (2011), "Automated Marketing Research Using Online Customer Reviews," *Journal of Marketing Research,* 48 (5), 881–94.

Marinova, Detelina, Sunil K. Singh, and Jagdip Singh (2018), "Frontline Problem-Solving Effectiveness: A Dynamic Analysis of Verbal and Nonverbal Cues," *Journal of Marketing Research*, 55 (2), 178–92.

Matz, Sandra and Oded Netzer (2017), "Using Big Data as a Window Into Consumers' Psychology," *Current Opinion in Behavioral Sciences*, 18, 7–12.

McDuff, Daniel and Jonah Berger (2020), "Why Do Some Advertisements Get Shared More Than Others?," *Journal of Advertising Research*, 60 (4), 370–80.

Melumad, Shiri, J. Jeffrey Inman, and Michel Tuan Pham (2019), "Selectively Emotional: How Smartphone Use Changes User-Generated Content," *Journal of Marketing Research*, 56 (2), 259–75.

Mosley, Buffy, David A. Schweidel, and Kunpeng Zhang (2024), "When Connection Turns to Anger: How Consumer–Brand Relationship and Crisis Type Moderate Language on Social Media," *Journal of Consumer Research*, 50 (5), 907–22.

Packard, Grant and Jonah Berger (2021), "How Concrete Language Shapes Customer Satisfaction," *Journal of Consumer Research*, 47 (5), 787–806.

Packard, Grant, Yang Li, and Jonah Berger (2023), "When Language Matters," *Journal of Consumer Research*, 51 (3), 634–53.

Rocklage, Matthew D., Jonah Berger, and Reihane Boghrati (2025), "The Trajectory of Confidence: Experience, Certainty, and Consumer Choice," *Journal of Marketing Research*, 63 (2), 364–82.

Rocklage, Matthew D. and Russell H. Fazio (2015), "The Evaluative Lexicon: Adjective Use as a Means of Assessing and Distinguishing Attitude Valence, Extremity, and Emotionality," *Journal of Experimental Social Psychology*, 56, 214–27.

Rocklage, Matthew D. and Russell H. Fazio (2020), "The Enhancing Versus Backfiring Effects of Positive Emotion in Consumer Reviews," *Journal of Marketing Research*, 57 (2), 332–52.

Schramla, Christopher (2025), "Automated Video Analytics in Marketing Research: A Systematic Literature Review and a Novel Multimodal Large Language Model Method," OSF Preprints (63nbc_v1), https://osf.io/download/682cebe1f8b3e01b3bf07f42/.

Schwenzow, Jasper, Jochen Hartmann, Amos Schikowsky, and Mark Heitmann (2021), "Understanding Videos at Scale: How to Extract Insights for Business Research," *Journal of Business Research*, 123, 367–79.

Umashankar, Nita, Kihyun Hannah Kim, and Thomas Reutterer (2023), "Understanding Customer Participation Dynamics: The Case of the Subscription Box," *Journal of Marketing*, 87 (5), 719–35.

Yang, Jeremy, Juanjuan Zhang, and Yuhan Zhang (2025), "Engagement That Sells: Influencer Video Advertising on TikTok," *Marketing Science*, 44 (2), 247–67.

Yin, Dezhi, Samuel D. Bond, and Han Zhang (2017), "Keep Your Cool or Let It Out: Nonlinear Effects of Expressed Arousal on Perceptions of Consumer Reviews," *Journal of Marketing Research*, 54 (3), 447–63.