

Visual-Textual Synergy: A Multimodal Analysis of Female Identity Construction in Xiaohongshu Digital Advertising

Hanwen Zhang^{1*}, Beiyi Guo¹, Lixin Gao¹

¹Northeastern University, Shenyang, China

*Corresponding Author. Email: 20220485@stu.neu.edu.cn

Abstract. This study investigates the construction of female identity through multimodal (visual-textual) resources in Xiaohongshu digital advertising, focusing on YSL's 1966 lipstick advertising video. Integrating Multimodal Discourse Analysis (MDA), Gender Performativity Theory, and Russell's Circumplex Model of Emotion, a mixed-methods experiment engaged 10 female participants and 30 valid questionnaires results with interview results to compare three video versions: image-only, text-only, and full-modality. Results demonstrated that the full-modality condition elicited the strongest emotional engagement, with significant correlations between pleasure, confidence, and reduced stress. The synergy of visuals and textual affirmations reinforced narratives of modern femininity characterized by autonomy and aspiration. Future research should include a larger sample size and multiple product focus, and further expand to cross-cultural contexts, non-binary identities, and auditory modalities to deepen understanding of multimodal identity construction. This study highlights the transformative potential of multimodal advertising in shaping social perceptions of gender and urges critical engagement with digital media's symbolic complexity.

Keywords: Multimodal discourse analysis, gender performativity, identity construction, emotional engagement

1. Introduction

In the digital age, advertising plays a key role in shaping consumer identity, particularly through the use of visual and textual elements. One key area is promotional videos on social media platforms, which combine multimodal resources such as images, text, and sound to construct identity. As advertising increasingly moves to digital spaces, such as the leading Chinese social media platform Xiaohongshu, the role of these multimodal resources in shaping consumer perceptions becomes more evident.

Previous scholarship on identity construction largely treats identity as a social accomplishment: individuals form and revise self-understandings through ongoing engagement with others and with culturally available meanings [1]. Within early cultural and media studies, researchers examined how identity is staged and circulated across media, highlighting the symbolic role of media in consolidating collective identifications [2]. Building on this trajectory, recent multimodal discourse analysis (MDA) has extended identity research to the metaphorical and semiotic shaping of female identity in particular [3].

Gender identity in advertising has long been a central concern in the literature. Instead of treating gender as a fixed inner property, most contemporary accounts frame it as something people do—an identity produced through repeated acts and culturally legible ways of behaving [4]. Within this logic, advertisements have traditionally cast women in familiar feminine scripts, supported by visual styling and verbal framing that highlight beauty, delicacy, emotional warmth, or quietness [5]. Yet this is no longer the only direction available. Increasingly, campaigns also depict women as decisive, competent, and self-possessed, a shift that resonates with changing public attitudes toward gendered roles and expectations.

Even so, an important gap remains. Existing work has not yet fully explained how multimodal resources in advertising—especially the coordination of image, text, sound, and platform-specific features—actively participate in producing particular versions of female identity. While traditional advertising has been studied extensively in terms of female imagery [6], far fewer studies address digital advertising contexts where visual and textual elements are tightly interwoven and rapidly consumed. To respond to this gap, the present study investigates how video advertisements integrate visual and verbal resources to construct specific female identities, using the YSL 1966 lipstick campaign on Xiaohongshu as a case. Drawing on a multimodal discourse analysis framework [7], the study examines how images, on-screen text, and other symbolic cues combine to generate meaning, shape consumer affect, and articulate identity positions. The research asks: (1) How do textual and visual elements work together to construct female identity in advertising? (2) What kinds of female identities are produced and communicated through these multimodal configurations?

By addressing these questions, this study helps to gain a deeper understanding of how digital advertising uses multimodal resources to shape consumer identity. By focusing on how images and text work together to create meaning, this study provides new insights into the role of digital advertising in influencing identity construction. As social media platforms continue to evolve, understanding the meaning of multimodal discourse in advertising is essential for academic research.

2. Theoretical framework

This study is anchored in three complementary perspectives: multimodal discourse analysis (MDA) [7], gender performativity theory [4], and the circumplex (emotion-cycle) model of affect [8]. Taken together, these approaches allow the analysis to move across meaning-making, identity performance, and audience response. From an MDA standpoint, the study examines how multiple semiotic resources—especially images, written text, and their patterned co-deployment—work in concert to build identity meanings. Butler’s account of gender performativity provides the conceptual tools to interpret how femininity is enacted, regulated, and signaled through these multimodal choices. Russell’s model of emotion, in turn, clarifies how such identity performances may provoke affective reactions that shape viewers’ interpretations and potential alignment with the female identity being promoted. Using this integrated framework, the study aims to trace the layered and sometimes subtle ways female identity is produced and circulated in digital advertising contexts.

2.1. Multimodal Discourse Analysis (MDA)

Multimodal discourse analysis, developed by Kress and van Leeuwen [7], offers a systematic way to understand how distinct semiotic modes—image, language, sound, movement, gesture, and layout—cooperate to generate meaning. Their visual grammar highlights three metafunctions: representational, interpersonal, and compositional.

The representational metafunction concerns how people, objects, and events are depicted as parts of social reality. In promotional videos, for instance, the way women are shown through posture,

styling, clothing, or spatial setting can reaffirm familiar gender scripts or rework them by introducing alternative identity cues.

The interpersonal metafunction concerns how an advertisement sets up a relationship among the producer, the viewer, and the people shown on screen. Gaze is a particularly clear case. When a woman in the ad looks straight at the audience, the video positions viewers as being directly addressed, almost pulled into interaction; when her gaze is averted, viewers are placed in a more detached, observational role. In this way, gaze helps determine whether the scene “demands” attention and engagement or simply “offers” an image to be looked at.

The compositional metafunction refers to how the multimodal components are organized so that some elements stand out and others recede. This involves choices about framing, salience, ordering, and spatial layout—basically, how the video arranges meaning in front of the viewer. Such design decisions guide interpretation by signaling what should be read as the focal point, what functions as supporting detail, and how different cues are meant to connect into an overall message.

Because digital cosmetic advertising typically relies on tightly coordinated visuals and short textual prompts, MDA is especially useful for showing how identity is shaped through the interaction of modes, and how these interactions produce meanings about femininity, power, and self-presentation [9].

2.2. Gender performativity theory

Judith Butler’s theory of gender performativity rejects the idea that gender is a natural or fixed property. Instead, gender is understood as an ongoing accomplishment—less something one is than something one repeatedly does within recognizable cultural norms [4]. Gender identity emerges through stylized repetition of actions, behaviors, and discursive forms, and becomes intelligible because those repetitions are socially regulated and culturally legible.

In advertising, this performative character of gender is often made visible. Campaigns stage femininity through multimodal choices: the body as a visual sign, the voice or caption as a discursive cue, and the overall narrative as a cultural script. Such resources can reproduce dominant performances of femininity, but they can also reshape them—for example, by pairing images of women with language that foregrounds autonomy or authority. This theoretical lens therefore supports the analysis of how female identity is enacted in promotional videos, and how particular gendered performances are constructed through coordinated visual and textual design.

2.3. Circumplex model of emotion

Russell’s circumplex model of emotion [8] treats affect as varying along two linked continua: arousal, which runs from low-energy calm to high-energy activation, and valence, which runs from negative to positive feelings. Thinking in these terms helps capture not just what viewers feel when watching an advertisement, but how strongly they feel it and whether that feeling pulls them toward or away from the message, especially in multimodal settings where meaning is built across modes.

In advertising, this emotional layer is not incidental—it is one of the main routes through which persuasion works. Visual sequences that portray women as composed, confident, or in control can push emotions toward positive valence and higher arousal. Short on-screen captions often intensify or steer these responses by naming the feeling or framing it as desirable and reachable. For example, a lipstick video may highlight a woman applying makeup with deliberateness and ease, while captions evoke empowerment or self-definition. Read together, the visual choreography and the verbal cues can foster emotions like enjoyment, excitement, or a sense of self-affirmation.

Focusing on these affective consequences matters because emotion is a bridge between representation and identification. What viewers feel while watching shapes how they interpret the

femininity being presented—whether it comes across as admirable, relatable, or authoritative—and those impressions can feed into both consumption tendencies and broader gendered self-understandings [10].

3. Methodology

This study adopts a mixed-methods design, bringing qualitative reading together with quantitative assessment. Using both approaches makes it possible to track overall patterns in how multimodal arrangements shape responses, while also preserving the finer-grained ways participants describe and make sense of the identities they encounter.

3.1. Experimental design

To test how visual and verbal resources work together in shaping female identity, the study employed a controlled within-subjects experiment. Participants watched three edited versions of the same YSL 1966 lipstick promotional video, each version altering the balance between image and text so that the contribution of each modality could be examined more directly. The three conditions were as follows:

Image-enhanced version: all written and spoken text (e.g., subtitles, slogans, ad copy) was removed. What participants saw was only the visual layer of the video, including the model's appearance, the makeup-application process, her gestures and movement, and the broader setting.

Text-enhanced version: the visual track was eliminated, leaving only the verbal channel. This condition retained elements such as voiceover (where applicable), written slogans, subtitles, and any other textual prompts.

Full-modality version: both visual and textual components were kept intact, reproducing the original multimodal advertisement.

Each participant viewed all three versions, with the order counterbalanced to reduce sequence or carryover effects. Using a within-subjects design also helped control for stable individual differences—such as prior cosmetic experience or baseline views on gender roles—so that observed changes could be more confidently attributed to modality configuration. The experiment therefore aimed to compare how different combinations of image and text influence (a) viewers' interpretations of the female identity promoted in the ad and (b) the emotional responses elicited during exposure.

3.2. Data collection

The data collection process was structured into two sequential phases: Participant recruitment and screening based on predefined demographic and behavioral criteria, and a controlled experimental procedure designed to capture both quantitative and qualitative responses to multimodal stimuli.

3.2.1. Participants

The study recruits 10 female participants, aged between 18 and 25, who are regular consumers of beauty and cosmetic products. Participants were selected based on the following criteria: first, they should use cosmetics regularly (daily or several times a week). Second, they should be familiar with advertisements on Little Red Book. Third, there was no professional or academic experience in related research to avoid bias.

3.2.2. Research procedure

The experiment consists of the following steps. Initially, the 10 participants will be asked to complete a demographic survey to collect information about their age, makeup usage habits, and prior exposure to makeup advertisements. Then, participants will be shown the three versions of the YSL1966

lipstick promotional video in the order of text-enhanced version, image-enhanced version, and full-modality version. After viewing each version, participants will be asked to complete a series of questionnaires. Using a Likert scale (1–5), participants will rate their emotional reactions to the video. Participants will also be asked to describe how they perceive the female identity represented in the video, with a focus on traits such as confidence, independence, elegance, or vulnerability. So, a total of 30 valid questionnaire results about participants' reaction to the three versions of the video were made. In a follow-up interview, participants will be asked to discuss their views on the interaction between text and image in shaping the identity of the woman in the video. The interview aims to understand whether the combination of text and image enhanced or diluted their perception of the female identity in the advertisement.

3.3. Data analysis

Data analysis is conducted using IBM SPSS Statistics 27, including descriptive statistics, correlation analysis, and paired samples t-tests. Descriptive statistics are used to calculate the mean and standard deviation for emotional responses (e.g., pleasure, confidence, excitement) for each video version. This provides an overview of how participants emotionally respond to the different conditions. Correlation analysis (Spearman's rho) examines the relationship between emotional responses and identity perception, assessing how emotions like excitement and confidence relate to the recognition of female identity traits (e.g., self-assurance, elegance). Finally, paired samples t-tests compare emotional responses and identity perceptions across the three video conditions (image-only, text-only, full modality) to determine if there are significant differences within participants.

4. Results

To examine whether emotional reactions are associated with identity perceptions across the three experimental conditions (text-enhanced, image-enhanced, and full-modality), this study applies Spearman's rho correlation analysis. The sections below report the key emotion–emotion correlations and discuss what they suggest about multimodal identity construction.

4.1. Correlations among emotions

As shown in Table 1 for the text-enhanced condition, pleasure is negatively related to stress ($\rho = -0.525$, two-tailed $p = 0.119$). Although this correlation does not meet the conventional significance criterion ($p < 0.05$), its magnitude indicates a noticeable downward trend: participants who report greater pleasure also tend to report lower stress. In practical terms, this pattern is consistent with the idea that pleasurable engagement with the text-focused advertisement may dampen feelings of pressure or tension, even if the present sample does not provide strong enough evidence to confirm the effect statistically.

A much clearer association appears between aspiration and perceived alignment with an ideal state. The correlation is strong and positive ($\rho = 0.817$, $p = 0.004$), indicating a statistically significant relationship. Participants with higher aspiration scores are more likely to feel that they are closer to their ideal selves. This result implies that aspiration is closely tied to how viewers evaluate the distance between who they are and who they want to be. One plausible interpretation is that aspirational orientations make ideal self-images more salient and actionable, which in turn increases perceived proximity to that ideal.

Most other correlations between the variables analyzed do not show significant relationships.

Table 1. Correlations among emotions in text-enhanced group

		Image Sensitivity	Pleasure	Aspiration	Confidence	Emotional Connection
Stress	Correlation Coefficient	-.323	-.525	.173	-.028	.272
	Sig.	.363	.119	.633	.939	.447
Alignment with Ideal State	Correlation Coefficient	.231	-.181	.817	.037	.187
	Sig.	.521	.617	.004	.919	.605

In table 2, which shows correlations among emotions in image-enhanced group, a positive correlation ($r = 0.570$, $p = 0.085$) is shown between feelings of pleasure and feelings of confidence. While the positive correlation between these two variables suggests that an increase in pleasure may lead to an increase in confidence, the p-value is close to 0.1 and does not reach the traditional level of significance (usually $p < 0.05$), which indicates that this correlation is at a borderline level of significance. Specifically, this result shows that when viewers feel more pleasure while watching the image version of the advertisement, they may simultaneously feel a greater sense of confidence.

The correlation between pleasure and stress is strong, negative, and statistically significant ($r = -0.761$, $p = 0.011$). In other words, participants who reported more pleasure while watching the advertisement tended to report less tension or stress; when stress increased, pleasure generally fell. Aside from this pairing, the remaining correlations among the measured variables were small and did not reach statistical significance.

According to Table 3, which illustrates correlations among emotions in full-modality group, a positive correlation was found between feelings of pleasure and feelings of aspiration ($r = 0.745$, $p = 0.013$), and a significant positive correlation is also found between feelings of pleasure and feelings of confidence ($r = 0.745$, $p = 0.013$). This indicates that when female images in advertisements show pleasurable and positive emotions, viewers tend to develop a sense of aspiration and increased self-confidence towards such images. There is a significant negative correlation between pleasure and stress ($r = -0.761$, $p = 0.002$), which means that pleasure is effective in reducing viewers' feelings of tension.

Table 2. Correlations among emotions in image-enhanced group

		Image Sensitivity	Aspiration	Confidence	Stress	Emotional Connection	Alignment with Ideal State
Pleasure	Correlation Coefficient	.382	.323	.570	.761	.295	-.081
	Sig.	.277	.363	.085	.011	.409	.823

The positive correlation between emotional connection and sense of confidence ($r = 0.611$, $p = 0.061$) shows that viewers' emotional connection with the female image in the advertisement may enhance their confidence with the image. The negative correlation between emotional connection and feelings of stress ($r = -0.799$, $p = 0.006$) suggests that participants with greater emotional resonance may feel less nervousness or stress while viewing the advertisement.

Additionally, the positive correlation between alignment with ideal state and feelings of aspiration ($r = 0.756$, $p = 0.024$) indicates that viewers' feelings of aspiration for a female image in an advertisement are enhanced when they perceive the image to be in alignment with ideal state.

Table 3. Correlations among emotions in full-modality group

		Aspiration	Confidence	Stress
Pleasure	Correlation Coefficient	.745	.745	-.849
	Sig.	.013	.013	.002
Emotional Connection	Correlation Coefficient	.611	.611	-.799
	Sig.	.061	.061	.006
Alignment with Ideal State	Correlation Coefficient	.700	.700	-.242
	Sig.	.024	.024	.501

4.2. Relations and effects of modalities

This section will dissect the statistical correlations between image and text modalities and their different impact on participants' perceptions of femininity and identities.

4.2.1. Relations of modalities

Table 4 shows the correlation between the two variables, image impact and text impact. There is a significant positive correlation between image impact and text impact with a correlation coefficient of 0.640 and a p-value of 0.046, indicating that viewers' separate evaluations of image and text are synchronized. When viewers perceived the image impact to be greater, they also tended to perceive the text impact to be greater, which may reflect the synergistic effect of image and text in the video.

Table 4. Correlations between impact of text and impact of images

		Text Impact
Image Impact	Correlation Coefficient	.640
	Sig.	.046

Table 5. Correlations between impact of text and impact of images

	Frequency	Percent	Valid Percent	Cumulative Percent
Slightly Enhanced	3	30.0	30.0	30.0
Fully Enhanced	7	70.0	70.0	100.0

From the data in Table 5, it can be concluded that all participants felt that the combination of image and text modality enhanced their understanding of the video, with the vast majority of participants (70%) believing that the combination of text and image has a “fully enhanced” impact on them, and 30% believing that it has a “slightly enhanced” impact. This suggests that the combination of image and text modalities was more compelling than a single modality and had a stronger impact on the majority of viewers, better motivating them to feel emotion or identify with the video.

4.2.2. Effects of modalities

Table 6 illustrates the statistical analysis of “image sensitivity” among different groups, including text, image and combination of text and image modalities. Specific data show that the average sensitivity of graphic combination is the highest (4.4000), followed by text (3.5000), and the average sensitivity of image is close to that of text (3.4000). For the plurality, both image and graphic combination were 4.00 and text was 3.00 (multiple plurality existed and the smallest value is shown). Overall, the combination of graphic and text forms performed optimally in terms of sensitivity, with

an overall higher sensitivity than a single modality of text or image, and a more concentrated distribution of data. Thus, modalities can influence participants' perception toward femininity. The combination of two modalities, image and text, can significantly increase participants' sensitivity to the image of women in the video compared to a single modality.

According to Table 7, it shows the statistical analysis of the different types of “conformity to the ideal state”, including text, image and graphic-text combination. Specific data show that participants in the graphic-text combination had the highest mean (2.4000), followed by the image (1.7000), and the text had the lowest mean (1.5000), for the perception that the identity of the woman in the video conformed to the ideal state. For the plurality, graphic combination was 3.00, image was 2.00, and text was 1.00 (multiple plurality existed and the smallest value is shown). It is concluded that the combined graphic and text format performed best in terms of participants' perception that the female identity in the video conformed to the ideal. This suggests that modality affects participants' judgment of whether the female identity in the video conforms to the ideal state. With the combination of modality, participants may have more positive emotions and aspire to the female identity in the video and perceive it as more consistent with the ideal state.

Table 6. Image sensitivity among three groups

	image sensitivity (text-enhanced group)	image sensitivity (image-enhanced group)	image sensitivity (full-modality group)
Mean	3.5000	3.4000	4.4000
Mode	3.00a	4.00	4.00

Table 7. Alignment with ideal state among three groups

	Alignment with Ideal State (text-enhanced group)	Alignment with Ideal State (image-enhanced group)	Alignment with Ideal State (full-modality group)
Mean	1.5000	1.7000	2.4000
Mode	1.00a	2.00	3.00

5. Discussion

Drawing on Kress and van Leeuwen's multimodal discourse analysis (MDA), Russell's emotional cycle model, and Butler's theory of gender performativity, this study reveals how images and texts work together to shape audiences' perceptions of femininity, emotional engagement, and ideal identity.

5.1. Multimodal synergy and identity construction

The results demonstrate that the full-modality condition (combining image and text) elicited the strongest emotional and cognitive responses, as evidenced by higher ratings in image sensitivity (Table 6) and alignment with the ideal state (Table 7). Applying Kress and van Leeuwen's MDA framework, the representational metafunction is evident in how the video's visuals (e.g., a confident model applying lipstick) and text (e.g., “Empower your inner beauty”) jointly reinforce a narrative of modern femininity characterized by autonomy and self-assurance. The interpersonal metafunction operates through the model's direct gaze and assertive posture, which, coupled with empowering language, establishes a relationship of equality and inspiration with the viewer [9]. Meanwhile, the compositional metafunction—seen in the harmonious layout of visual and textual elements—enhances coherence, ensuring that the combined modalities amplify the message of female empowerment [7].

These findings align with Butler's notion of gender performativity, where femininity is not innate but enacted through repeated symbolic practices. The advertisement performs a specific female identity (confident, aspirational), and aligned with contemporary ideals through the interplay of visual

cues (e.g., bold makeup application) and textual rhetoric. This identity is further naturalized through the multimodal synergy, as viewers internalize the performance as a socially validated norm [5].

5.2. Emotional engagement and identity internalization

Drawing on Russell's Circumplex Model of Emotion, this section explains how multimodal configurations in the advertisement generate affective states that, in turn, shape identity construction. In the full-modality condition, pleasure, aspiration, and confidence show significant positive intercorrelations (see Table 3), a pattern consistent with high arousal and positive valence. Put differently, when lively visuals of lipstick application (arousal-enhancing cues) are paired with aspirational wording (valence-reinforcing cues), viewers are more likely to experience excitement and self-assurance and to connect those feelings to the female identity being promoted [10]. Alongside this, the strong negative relationship between pleasure and stress ($r = -0.761$) suggests that a coherent multimodal message can ease tension or psychological conflict, making it easier for participants to accept the identity narrative without resistance [6].

By comparison, responses in the two single-modality conditions were weaker and less consistent. In the image-enhanced version, pleasure showed only a marginal link with confidence ($r = 0.570$, $p = 0.085$), while in the text-enhanced version most emotional associations were non-significant (Table 1). Taken together, these patterns highlight the complementarity of modes: visuals tend to trigger fast, embodied feelings, whereas text helps specify what those feelings are about by attaching them to particular identity attributes (for instance, "independence"). When image and text operate in tandem, they produce a more coherent and convincing identity narrative—one that viewers can both experience emotionally and interpret cognitively [3].

5.3. Interactivity of modalities and audience perception

The results indicate a strong synergistic relationship between visual and textual modes. As Table 4 shows, perceived image impact correlates significantly and positively with perceived text impact ($r = 0.640$, $p = 0.046$), which implies that viewers generally experience these modalities as reinforcing each other rather than as separate streams. In concrete terms, a slogan like "Empower your inner beauty" carries more persuasive weight when it is paired with visuals of a poised and self-assured model. At the same time, the model's confidence is not left as a vague impression; the accompanying text gives it a clearer cultural and ideological frame. This matches the core claim of MDA that meaning arises through cross-modal interaction, not through isolated semiotic channels [7].

The interview data after the experiment support this reading. Many participants said that image and text together produced a stronger semiotic "push," helping them grasp the intended female identity more directly. One participant noted that the visuals first caught her attention, but the text prompted her to think more carefully about what confidence and femininity mean in the present moment. Such reflections are consistent with Kress and van Leeuwen's argument that multimodal coherence enhances ideological persuasion because it engages viewers emotionally while also guiding interpretation cognitively.

Although the study focused on visual and textual resources, several interviewees also pointed to the role of sound—especially voiceover delivery and background music. They described how upbeat music and a confident vocal tone made the model feel both more relatable and more aspirational. This suggests that sound, as an additional semiotic resource, can deepen emotional resonance and increase the salience of the promoted identity. From a performativity angle, auditory features such as tone, rhythm, and pacing may further ritualize gender performances by anchoring them in the viewer's sensory and affective experience, thereby strengthening identification at an embodied level.

Overall, participants tended to describe the female identity presented in the advertisement as confident, independent, and ambitious. Crucially, they traced these impressions to the combined work of modes: visual cues established the affective and behavioral texture of modern femininity, while textual affirmations named and stabilized its ideological direction. Through this multimodal alignment, the advertisement framed femininity as autonomous, goal-oriented, and consistent with contemporary empowerment ideals, repeatedly foregrounding traits such as confidence and ambition through the interaction of image and text.

5.4. Limitations and future directions

Although the study yields useful insights into multimodal identity construction, several limitations should be acknowledged. First, the small sample size ($N = 10$) restricts the extent to which the results can be generalized beyond this participant group. Second, because the analysis centers on a single product (YSL 1966 lipstick) and one platform environment (Xiaohongshu), the findings may not transfer straightforwardly to other advertising genres, brands, or digital ecosystems [1]. These constraints also point to clear directions for future work. Subsequent studies could test a wider range of products and media platforms, include representations of male and non-binary identities, and examine how multimodal identity cues function across different cultural settings [2]. In addition, participants repeatedly mentioned the influence of sound in shaping their reactions, suggesting an underexplored but important modality. Future research that explicitly investigates how auditory features—such as voice quality, rhythm, and background music—interact with visual and textual resources would help build a more complete account of how multimodal discourse operates holistically in identity formation [11].

6. Conclusion

Using the YSL 1966 lipstick campaign on Xiaohongshu as a case, this study examined how multimodal resources collaborate to construct female identity in short-form promotional video. By bringing together multimodal discourse analysis (MDA) [7], gender performativity theory [4], and the circumplex (emotion-cycle) model of emotion [8], the analysis shows that identity meanings in digital advertising are produced through the combined operation of visual design, textual framing, and the emotional states they elicit.

Across all three experimental conditions, the full-modality video—where visuals and text were presented in tandem—produced the most robust emotional and interpretive responses. Participants exposed to this version reported greater pleasure, desirability, and confidence, implying that when modalities are aligned, the advertised identity becomes easier to experience affectively, to interpret, and ultimately to accept. Read through an MDA lens, these outcomes point to the joint work of representational design (the type of woman being depicted), interpersonal positioning (the relationship the video builds with viewers), and compositional structuring (how attention and meaning are organized). Together, these metafunctions foreground a contemporary femininity marked by autonomy, composure, and self-directed agency. Butler's concept of gender performativity [4] sharpens this interpretation: the advertisement is not merely showing femininity as a quality, but orchestrating it as a repeatable performance that feels “natural” and socially valued because it is stabilized through multimodal repetition and coherence.

At the same time, the study should be read within its empirical boundaries. The sample was small ($N = 10$), and the analysis drew on one product and one platform context, which limits how far the findings can be extended to other genres of advertising, cultural settings, or audience groups. Future work could address this by expanding the corpus across brands and platforms, testing audiences in different cultural environments, and adding further modes—especially sound and voice—to better

capture the full ecology of multimodal identity production. It would also be important to examine how these identity dynamics intersect with masculine and non-binary positions, rather than leaving femininity as the only analytic focus.

Taken together, the results reaffirm that multimodal advertising has substantial power to shape how gendered identities are envisioned, felt, and taken up in everyday digital consumption. They also call for ongoing critical attention to the ideological and semiotic mechanisms through which digital campaigns make particular identities appear persuasive, desirable, and ultimately “normal.”

References

- [1] Cerulo, K. A. (1997). Identity construction: New issues, new directions. *Annual Review of Sociology*, 23(1), 385-409. <https://doi.org/10.1146/annurev.soc.23.1.385>
- [2] Sa'ar, A., & Simchai, D. (2023). Millennial femininity and the harmonious state of mind. *Current Sociology*, 71(1), 114-132. <https://doi.org/10.1177/00113921221101967>
- [3] Zhao, X., & Xie, Y. (2020). Multimodal metaphorical construction of female images in perfume advertisements. *Journal of Zhejiang International Studies University*, 5, 31-39.
- [4] Butler, J. (1990). *Gender trouble: Feminism and the subversion of identity*. Routledge.
- [5] Gill, R. (2007). *Gender and the media*. Polity Press.
- [6] Rose, S. (2001). New age women: Spearheading the movement? *Women's Studies: An Interdisciplinary Journal*, 30(3), 329-350. <https://doi.org/10.1080/00497878.2001.10435210>
- [7] Kress, G., & van Leeuwen, T. (2006). *Reading images: The grammar of visual design* (2nd ed.). Routledge.
- [8] Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178. <https://doi.org/10.1037/h0077714>
- [9] Jewitt, C. (2014). *The Routledge handbook of multimodal analysis* (2nd ed.). Routledge.
- [10] Schmitt, B. H. (2011). *The consumer psychology of brands*. Routledge.
- [11] Goodwin, C. (2000). Action and embodiment within situated human interaction. *Journal of Pragmatics*, 32(10), 1489–1522. [https://doi.org/10.1016/S0378-2166\(99\)00096-X](https://doi.org/10.1016/S0378-2166(99)00096-X)