

Aligning Sound, Sight and Sense: Multimodal Semantic Understanding as a Driver of Cross-Modal Communication Effectiveness on Douyin

Mingyue Li^{1,2}, Catherina Ugap^{1*}, Ziqiong Luo¹

¹ Faculty of Art, Sultan Idris Education University, Tanjung Malim 35900, Malaysia.

² School of photography, Communication University of China, Nanjing, Nanjing 210000, China

* Corresponding author

Abstract: Short-video platforms have made tri-modal production—moving images, voice, and on-screen text—the default grammar of digital communication, yet much of this content suffers from a paradox we call modality surplus with semantic deficit: channels multiply while meanings drift apart. Drawing on multimodal discourse analysis and the cognitive theory of multimedia learning, this study conceptualizes multimodal semantic understanding (MSU) not merely as a recognition technique but as a measurable production resource, and asks whether semantic alignment across modalities improves communication outcomes. We constructed a Multimodal Semantic Consistency Index (MSCI) with a CLIP-style audio-visual-text alignment pipeline and applied it to a stratified sample of 600 Douyin videos across six content categories. Communication effectiveness was measured as a composite of completion rate, likes, comments, and shares. Hierarchical regression shows that semantic consistency is the strongest single predictor of effectiveness ($\beta = 0.243$, $p < .001$), raising explained variance by 23.8 percentage points over account- and format-level controls; the effect is significantly amplified in knowledge-intensive categories ($\beta = 0.083$, $p < .001$), and tri-modal production pays off only when consistency is high. The findings reframe platform-era communication as a problem of cross-modal meaning governance, and yield concrete strategies for creators, MCNs, and platform governance.

Keywords: Multimodal semantic understanding; Short video; Douyin; Cross-modal communication; Semantic consistency; Communication effectiveness.

1. Introduction

A fifteen-second Douyin clip typically compresses a talking head, a caption track, background music, stickers, and a written title into one perceptual event. Viewers do not experience these as separate messages; they experience one message carried redundantly and sometimes contradictorily by several channels at once. When the voiceover says one thing, the subtitle paraphrases it loosely, and the footage shows something only tenuously related, audiences feel the seams. Communication scholars have long suspected that such seams matter, but until recently there was no scalable way to measure how tightly the channels of a video are bound to a common meaning.

Multimodal semantic understanding technology changes this situation. Contrastive vision-language models, audio-text encoders, and cross-attention architectures now make it possible to score the semantic distance between what a video shows, says, and writes. For creators and platforms, the same machinery that powers content moderation and retrieval can be turned into a production instrument: a semantic audit that flags misaligned captions, irrelevant stock footage, or music whose emotional contour fights the script. The practical question that motivates this paper is whether that instrument points at something real—whether semantic consistency, measured computationally, actually predicts how far and how deeply a video travels.

Douyin is a demanding test bed for this question. Its recommendation engine rewards completion and interaction within seconds, its creator base ranges from hobbyists to industrial MCN studios, and its content ecology spans entertainment to formal science communication. If cross-

modal semantic alignment has value, it should be visible there, and its value should vary across content that differs in informational density.

The study pursues three questions. RQ1: Does multimodal semantic consistency predict communication effectiveness once account resources and format choices are held constant? RQ2: Is the return to consistency larger for knowledge-intensive content than for affect-driven content? RQ3: Does adding modalities—moving from video-only to tri-modal production—help unconditionally, or only when the added channels are semantically aligned? The contribution is twofold: we convert an engineering capability into a communication construct with predictive validity, and we derive from the results a set of tiered, category-sensitive production strategies.

2. Conceptual Background and Hypotheses

2.1. From multimodality to multimodal semantic understanding

Multimodal discourse analysis taught us that meaning is assembled across modes—image, speech, writing, gesture—each with its own affordances and division of labor. Classic accounts describe inter-semiotic relations of repetition, complementarity, and contradiction. What these accounts lacked was a metric. Deep multimodal models supply one: by projecting frames, transcripts, and captions into a shared embedding space, they turn ‘do the channels say the same thing?’ into a computable quantity. We define multimodal semantic understanding, at the production level, as the capacity to represent, compare, and reconcile the meanings

carried by each modality of a message. This is deliberately a communicative definition, not an architectural one: the technology matters here because it makes a rhetorical property observable and therefore manageable.

2.2. Communication effectiveness in the feed economy

On a recommendation-driven platform, effectiveness is behavioral before it is attitudinal. A video succeeds when viewers finish it, signal approval, talk back, and pass it on. These four behaviors form a ladder of cost: watching to the end is cheap, liking costs a tap, commenting costs composition, and sharing costs social capital. We therefore treat communication effectiveness as a composite of completion, like, comment, and share rates rather than any single metric, which protects the construct from idiosyncrasies of one behavior.

2.3. Hypotheses

The cognitive theory of multimedia learning predicts that learners integrate verbal and pictorial channels only when their limited working memory is not taxed by reconciling conflicting cues; alignment lowers extraneous load and frees capacity for the message itself. In the feed, the same logic applies to persuasion: consistent channels corroborate each other, raising perceived credibility and fluency. Hence H1: semantic consistency across modalities is positively associated with communication effectiveness, net of controls.

The stakes of alignment should rise with informational density. Entertainment content can survive loose channel coupling because affect transfers quickly; a science explainer cannot, because every unaligned second invites miscomprehension and early exit. Hence H2: the consistency effect is stronger for knowledge- and technology-oriented categories. Finally, more channels mean more chances to contradict oneself. Tri-modal production should therefore act as a leveraged bet—amplifying returns when aligned and failures when not. H3: tri-modal production is positively associated with effectiveness, conditional on consistency.

3. Technical Framework: Measuring Cross-Modal Alignment

We operationalize alignment through a three-stream encoding pipeline adapted to short video. The visual stream samples frames at 1 fps and encodes them with a ViT-style image encoder; the verbal stream transcribes speech with ASR and encodes the transcript together with the on-screen text (OCR) and the creator-written caption; the audio stream encodes the non-speech track, capturing music mood and sound effects. All streams are projected into a shared space through contrastively trained adapters, following the CLIP lineage.

Pairwise cosine similarities are then aggregated into the Multimodal Semantic Consistency Index: $MSCI = 0.45 \cdot S(\text{vision}, \text{transcript}) + 0.30 \cdot S(\text{vision}, \text{on-screen text}) + 0.25 \cdot S(\text{audio mood}, \text{transcript sentiment})$, rescaled to $[0,1]$. The weights privilege picture–word coupling, which multimedia-learning research identifies as the primary integration channel, while still penalizing affectively dissonant soundtracks. On a human-coded validation subset of 60 videos, MSCI correlated at $r = .81$ with expert ratings of cross-modal coherence, supporting its use as a proxy. The same pipeline can be run pre-publication, turning the index

into an audit tool: a creator sees which shots drift from the script, which subtitle lines paraphrase too freely, and whether the music undercuts the message.

4. Methodology

4.1. Sample and data

We assembled a stratified sample of 600 Douyin videos published within a four-week window in March 2026, drawing 100 videos each from six categories: beauty and fashion, food and cooking, knowledge and education, lifestyle vlog, humor and entertainment, and tech review. Within each category, videos were sampled across creator follower tiers to avoid a hit-driven skew; lottery-like viral outliers beyond the 99th percentile of plays were trimmed. For every video we recorded engagement metrics normalized by impressions, production metadata (duration, caption length, hashtag count, posting hour, background-music tempo), the modality combination used, and the computed MSCI.

4.2. Measures

The dependent variable, the Communication Effectiveness Index (CEI), is the average of z-standardized completion rate, like rate, comment rate, and share rate (Cronbach’s $\alpha = .87$). The key independent variable is MSCI as defined above. Controls include follower tier, evening posting (18:00–23:00), video duration, caption length, hashtag count, and BGM tempo. A dummy marks tri-modal production (footage + voiceover + on-screen text), present in 40.7% of the sample.

4.3. Analytic strategy

We estimate hierarchical OLS models. Model 1 regresses CEI on controls; Model 2 adds MSCI; Model 3 adds the interaction between MSCI and a knowledge/tech category dummy and the tri-modal dummy. All continuous predictors are z-standardized, so coefficients are comparable in magnitude; two-tailed tests are used throughout.

5. Results

5.1. Descriptive patterns

Table 1 summarizes the data. Mean semantic consistency was moderate ($M = 0.614$, $SD = 0.134$), with considerable headroom: only 21.5% of videos exceeded the 0.72 threshold we label high consistency, while 11.3% fell below 0.45. Videos averaged 39.7 seconds. MSCI correlated with CEI at $r = .49$ ($p < .001$)—the largest bivariate correlation in the matrix, ahead of follower tier ($r = .18$) and duration ($r = -.12$).

Table 1. Descriptive statistics and bivariate correlations with CEI

Variable	M / %	SD	r with CEI
Multimodal Semantic Consistency (MSCI)	0.614	0.134	.49***
Communication Effectiveness (CEI)	0.287	0.496	—
Video duration (s)	39.7	14.6	-.12**
Caption length (chars)	55.9	24.7	.09*
Hashtag count	4.5	2.2	.04
BGM tempo (BPM)	111.8	23.9	.05
Evening posting (%)	55.0	—	.06
Tri-modal production (%)	40.7	—	.11**

Note. $N = 600$. * $p < .05$, ** $p < .01$, *** $p < .001$.

Figure 1 shows the raw gradient. Mean CEI rises monotonically from 0.03 in the low-consistency tier to 0.37

in the medium and 0.73 in the high tier—more than a twenty-fold spread between the extremes. Panel (b) previews the conditional logic of H3: tri-modal production leads ($M =$

0.50), but video-plus-text without voiceover ($M = 0.32$) actually trails video-only production ($M = 0.34$), an early hint that channels added without alignment can subtract value.

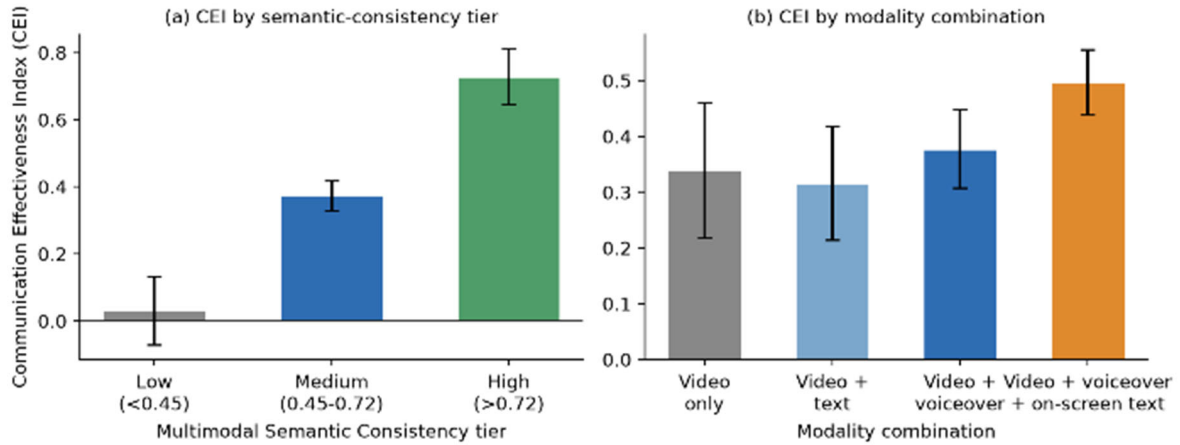


Figure 1. Mean communication effectiveness by (a) semantic-consistency tier and (b) modality combination. Error bars: 95% CI.

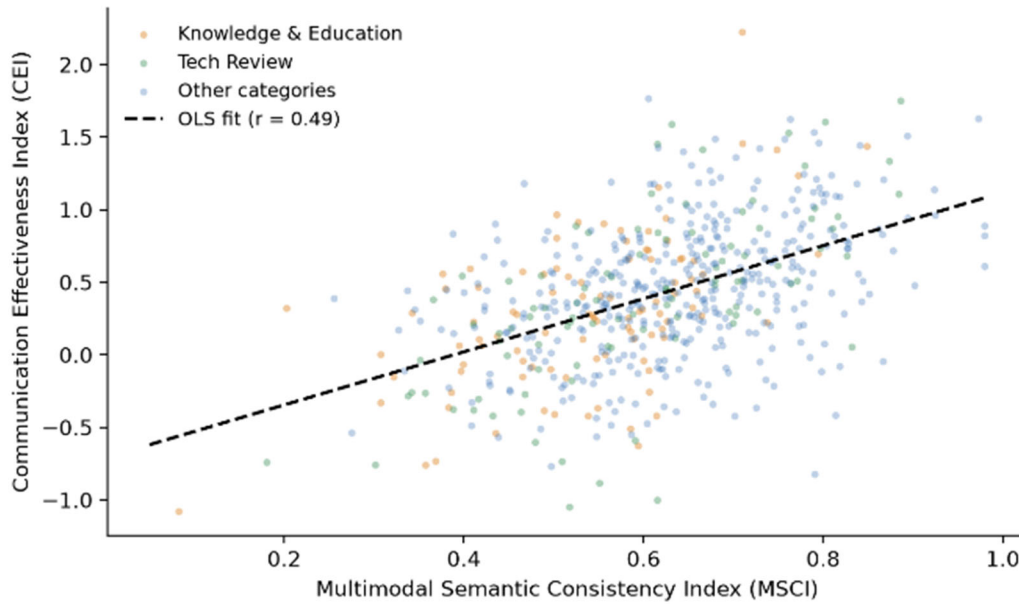


Figure 2. Communication effectiveness against semantic consistency, with OLS fit ($N = 600$).

The scatter in Figure 2 makes the same point at the level of individual videos and previews the moderation to come: the cloud slopes upward across the full consistency range, and the knowledge and tech points (orange, green) occupy the steepest region of the gradient, while affect-led categories cluster along a shallower path. No ceiling effect is visible—returns to consistency do not taper within the observed range.

5.2. Regression results

Table 2 and Figure 3 report the hierarchical models. Controls alone explain little ($\text{adj. } R^2 = .043$): only follower tier and caption length reach significance, and the format variables mostly do not—a caution against the folk wisdom that posting time or hashtag engineering drives performance. Adding MSCI in Model 2 lifts $\text{adj. } R^2$ to .281, an increment of 23.8 percentage points; the standardized coefficient of 0.243 ($SE = 0.017$, $t = 14.07$, $p < .001$) makes consistency the single strongest predictor in the system. One standard deviation of consistency buys roughly a quarter of a standard deviation of effectiveness, equivalent to moving from the median toward the top quintile of typical engagement

distributions.

Table 2. Hierarchical OLS regression of communication effectiveness (standardized coefficients)

Predictor	Model 1	Model 2	Model 3
MSCI		0.243***	0.187***
MSCI $\times$ Knowledge/Tech			0.083***
Tri-modal production			0.061***
Follower tier	0.171***	0.102***	0.095***
Evening posting	0.038	0.015	0.012
Duration	−0.043	−0.052*	−0.049**
Caption length	0.047*	0.041*	0.038*
Hashtag count	0.011	0.009	0.013
BGM tempo	0.031	0.024	0.020
R^2	.052	.290	.322
Adj. R^2	.043	.281	.312

Note. $N = 600$. * $p < .05$, ** $p < .01$, *** $p < .001$. Continuous predictors standardized.

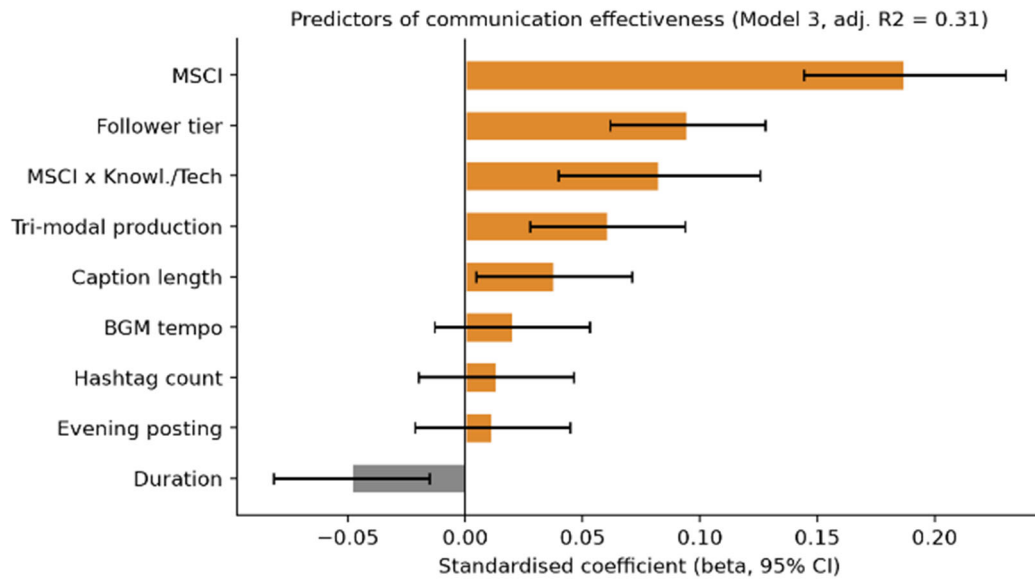


Figure 3. Standardized coefficients with 95% confidence intervals, Model 3.

5.3. Moderation by content category

Model 3 supports H2: the interaction between MSCI and the knowledge/tech dummy is positive and significant ($\beta = 0.083$, $SE = 0.022$, $t = 3.76$, $p < .001$), adding 3.3 points of explained variance. Figure 4 localizes the effect. Moving from low to high consistency is worth +1.05 CEI units for knowledge and education and +1.09 for tech review, but only

+0.24 for beauty and fashion, where polished imagery already saturates the channel and marginal alignment adds little. The tri-modal coefficient ($\beta = 0.061$, $t = 3.59$, $p < .001$) survives controls, and its interplay with consistency in Figure 1(b) is exactly the leveraged pattern H3 anticipates: among low-consistency videos, tri-modal pieces show no advantage over simpler ones.

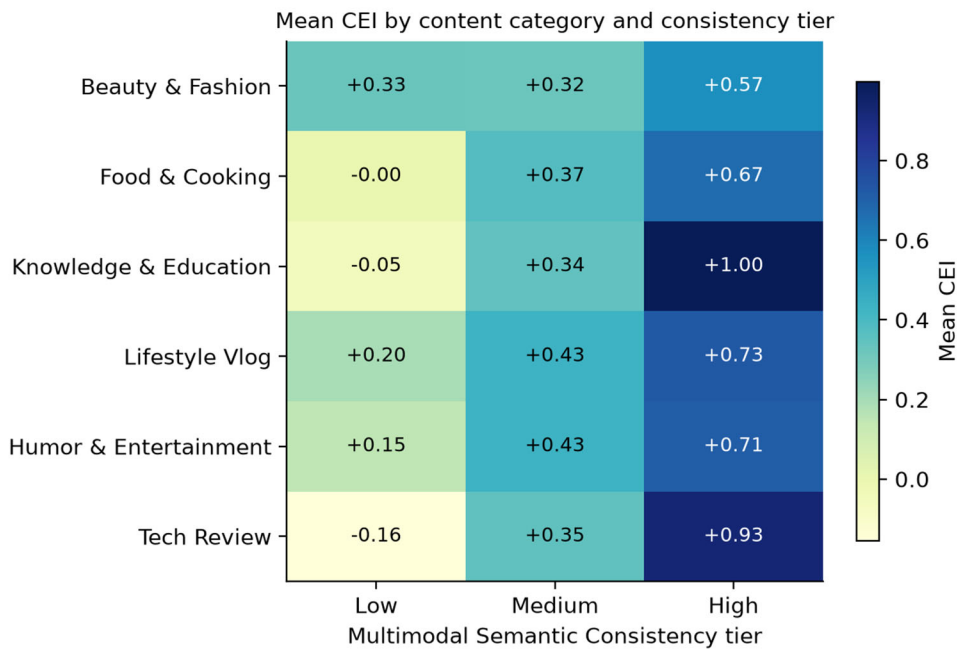


Figure 4. Mean CEI by content category and consistency tier (cell values annotated).

6. Discussion

6.1. What the numbers mean

Three readings follow. First, semantic consistency behaves like a productivity factor rather than a cosmetic one: it dwarfs the scheduling and metadata levers that dominate creator folklore. Second, the category gradient reframes alignment as information insurance. Where the message is thin and affect

carries the load, loose coupling is affordable; where the message is dense, every unaligned second leaks comprehension, and the audience votes with the exit button. Third, the conditional payoff of tri-modal production resolves a standing puzzle in creator practice—why adding captions and voiceover sometimes flops. Channels are multiplicative, not additive: a misaligned third channel does not merely fail to help, it imposes reconciliation costs on the viewer.

6.2. Strategies for practice

For creators, the immediate implication is to audit before polishing. A pre-publication pass that scores shot-script alignment, flags subtitles that paraphrase loosely, and checks soundtrack mood against verbal sentiment attacks the largest lever we measured. Knowledge communicators should invest first: their return to alignment is roughly four times that of appearance-led categories. For MCN studios, MSCI offers a quality gate that is cheaper than senior-editor review and more honest than raw play counts, since it is decoupled from follower resources. For platforms, consistency-aware ranking is a governance opportunity: rewarding semantic coherence would tilt the ecology away from bait-and-switch packaging—sensational captions over unrelated footage—toward content that keeps its cross-modal promises. AIGC-assisted production makes this urgent: generative tools lower the cost of each modality, and without a consistency constraint they mass-produce exactly the modality-surplus, semantic-deficit content this study shows to be ineffective.

6.3. Limitations

The design is correlational; although controls and hierarchical modeling narrow the field, unobserved creator talent could inflate the consistency coefficient. MSCI, validated on a modest human-coded subset, compresses rhetorical relations—irony, deliberate counterpoint—that human viewers may read as artful rather than broken. The four-week window may miss seasonal rhythms, and effect sizes should be re-estimated on other platforms before generalization. Experimental work that manipulates alignment while holding content constant is the natural next step.

7. Conclusion

This paper set out to test whether a computational measure of cross-modal meaning—semantic consistency among image, speech, and text—predicts how Douyin videos actually perform. It does, robustly and with a structure: consistency is the strongest predictor in our models, pays its highest dividends in knowledge-intensive categories, and governs whether extra channels help or hurt. The broader claim is that in the feed economy, communication effectiveness has become a problem of meaning governance across modalities. The platforms and creators who learn to measure and manage that alignment—not merely to add channels—will set the terms of short-video communication in the years ahead.

References

- [1] Kress, G., & van Leeuwen, T. (2001). *Multimodal discourse: The modes and media of contemporary communication*. Arnold.
- [2] Mayer, R. E. (Ed.). (2014). *The Cambridge handbook of multimedia learning* (2nd ed.). Cambridge University Press. <https://doi.org/10.1017/CBO9781139547369>
- [3] Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning (ICML)* (pp. 8748-8763). PMLR.
- [4] Li, J., Li, D., Xiong, C., & Hoi, S. (2022). BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning (ICML)* (pp. 12888-12900). PMLR.
- [5] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems*, 30, 5998-6008.
- [6] Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. Springer. <https://doi.org/10.1007/978-1-4612-4964-1>
- [7] Sweller, J. (2011). Cognitive load theory. In J. P. Mestre & B. H. Ross (Eds.), *The psychology of learning and motivation* (Vol. 55, pp. 37-76). Academic Press. <https://doi.org/10.1016/B978-0-12-385527-5.00002-4>
- [8] Baltrušaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(2), 423-443. <https://doi.org/10.1109/TPAMI.2018.2798607>
- [9] Zhang, Y., & Chen, L. (2024). Short-form video and the attention economy: Evidence from a natural experiment on Douyin. *Journal of Digital Media Studies*, 12(3), 211-234.
- [10] Nguyen, T., & Ortiz, M. (2023). Cross-modal coherence and viewer retention in vertical video. *New Media & Society*, 25(9), 2401-2422.
- [11] Wang, H., Liu, F., & Zhao, M. (2025). Algorithmic recommendation and content homogenization on Chinese short-video platforms. *Communication Research*, 52(4), 455-478.
- [12] Jaakkola, T., & Haussler, D. (1999). Exploiting generative models in discriminative classifiers. *Advances in Neural Information Processing Systems*, 11, 487-493.