Incorporating Semantic Visual Content into Click-Through Rate Prediction for Video Advertisements

Yoshiaki Tanabe*, Shuntaro Masuda*, Gakumatsu Ryu†, Naoto Tanji†, Hiroyuki Seshime†,
Ling Xiao*, and Toshihiko Yamasaki*
* The University of Tokyo, Japan
{tanabe, masuda, ling, yamasaki}@cvm.t.u-tokyo.ac.jp
† Septeni Japan, Inc., Japan
{gakumatsu.ryu, naoto.tanji, h_seshime}@septeni.co.jp

***Abstract*—This study presents a method for predicting the click-through rate (CTR) of video advertisements by leveraging high-level semantic content. While conventional CTR prediction models rely primarily on metadata such as ad categories or user behavior logs, our approach explicitly incorporates the semantic content of advertisements. We prompt GPT-4o to generate structured natural language descriptions of video scenes, which are then encoded into machine-interpretable semantic representations. To identify semantic features that reflect CTR drivers specific to the given dataset, we employ in-context learning with curated examples of high- and low-performing ads. The resulting interpretable features are selectively included as input features in the prediction model. Experimental results demonstrate that incorporating these dataset-specific semantic features reduces the mean squared error (MSE) by up to 14.02% compared to a baseline model. A case study further highlights that the extracted content not only improves predictive performance but also enhances model interpretability.**

I. Introduction

With the rise of short-form video platforms such as TikTok, the global market for video advertising has expanded rapidly [1]. As users are exposed to an overwhelming amount of content, their attention has become increasingly difficult to capture and maintain [2]. This shift highlights the growing need for advertisements to be both eye-catching and engaging. To improve ad effectiveness, it is essential to understand what types of content contribute to making an ad more attractive. In particular, identifying which high-level semantic features influence user engagement, through methods that are both quantitative and reproducible, has emerged as a key research challenge.

While click-through rate (CTR) prediction is a well-established task in computational advertising [3], conventional approaches typically rely on user metadata such as viewing history or user profiles, as well as low-level visual features extracted from pretrained convolutional neural networks such as ResNet50 [4], [5]. These features capture perceptual properties such as color distribution, edge patterns, features, and object presence, but they do not explicitly represent higher-level semantic features such as narrative structure or emotional appeal. Modeling such high-level, human-interpretable semantic factors remains challenging due to their abstract and context-dependent nature.

Foundation models are large-scale pretrained architectures that demonstrate strong generalization and reasoning capabilities across a wide range of tasks, including vision-language understanding. These models can leverage not only perceptual inputs but also extensive knowledge, allowing them to interpret content in contextually rich and semantically meaningful ways. In particular, Lin et al. [6] highlight the potential of leveraging foundation models' reasoning abilities and stored world knowledge to enhance the analysis of advertising effectiveness. Unlike conventional methods that treat ad content as unstructured pixels or rely solely on metadata, foundation models can generate structured natural language representations that better align with human understanding of an ad's semantic content.

In this study, we propose a CTR prediction framework that incorporates high-level semantic features using a multimodal foundation model. Specifically, we generate natural language descriptions of video content and extract semantic features that are representative of ad effectiveness. These features are then numerically embedded and integrated into a multimodal regression pipeline. Our approach differs from conventional methods by explicitly modeling human-interpretable semantic content. It enables both performance gains in CTR prediction and improved interpretability, thereby offering actionable insights for ad creators. Furthermore, by utilizing in-context learning, we allow the model to adaptively extract dataset-specific semantic drivers, improving robustness and generalizability. Our key contributions are as follows:

- We introduce a method for extracting effective high-level semantic features from video ads by leveraging in-context learning with a multimodal foundation model.
- We integrate these semantic features into a multimodal CTR prediction model, combining text, metadata, and visual embeddings.
- We empirically demonstrate that semantic features improve prediction performance and enables interpretable ablation analysis of ad effectiveness.

979-8-3315-7206-8/25/$31.00

II. Related Works

A. CTR Prediction with Foundation Models

Click-through rate (CTR) is a widely used metric to quantify consumer responses to advertisements [3]. Defined as the ratio of the number of clicks to the number of impressions, CTR prediction has attracted significant research attention, especially with the advancement of machine learning techniques [7]–[9]. While earlier work primarily focused on static image ads, recent studies have started exploring CTR prediction for video advertisements. He et al. [4] propose HyperCTR, a hypergraph neural network framework that models time-aware user-item interactions with multimodal features to learn modality-specific representations. Ikeda et al. [5] present a multimodal prediction framework that integrates video, text, and structured metadata with tailored normalization and regularization to mitigate overfitting on limited training data.

With the rise of large-scale pretraining and Transformer-based architectures [10], foundation models capable of generalizing across diverse tasks have gained popularity. Large language models (LLMs), a representative category, acquire advanced reasoning capabilities and broad world knowledge through pretraining on massive text corpora. For example, Brown et al. [11] demonstrated that GPT-3 performs well on various natural language reasoning and generation tasks without fine-tuning. Similarly, Petroni et al. [12] showed that Bidirectional Encoder Representations from Transformers (BERT) [13] can answer fact-based questions with high accuracy without task-specific training. The use of foundation models in CTR prediction is an emerging area of interest, though most existing approaches focus on processing structured metadata. For instance, Geng et al. [14] proposed applying partial layers of an LLM to long-term user behavior histories, generating refined user embeddings for CTR prediction. Lin et al. [15] introduced a framework called ClickPrompt, which integrates collaborative signals from conventional CTR models with semantic signals generated by pretrained language models. Their approach outperforms models based solely on metadata or language models.

However, relatively little research has examined how the high-level semantic features of video advertisements affect CTR. If the semantic features influencing consumer responses can be extracted in a human-interpretable form and analyzed quantitatively, they could offer a systematic basis for designing more effective advertisements. Thus, incorporating visual semantic features into CTR prediction for video ads presents a promising and largely unexplored research direction.

B. Video Understanding with Foundation Models

The application of foundation models to video understanding has received growing attention in recent years. Ko et al. [16] segmented videos into scenes and applied BLIP-2, an image captioning model, to generate captions for each scene, which were then concatenated and used as part of a prompt. Ataallah et al. [17] captioned individual frames using EVA-CLIP and fed them into a language model through a linear projection layer.

Meanwhile, foundation models such as Gemini [18] and GPT-4o [19] have recently been extended to handle multimodal inputs and outputs, including texts, images, and audio. Unlike traditional approaches that require explicit conversion of video frames into textual descriptions, these multimodal models can directly process visual content and infer its semantic features. In this study, we utilize GPT-4o to generate natural language descriptions of the semantic content of video advertisements in natural language and use the resulting textual representations as input features for CTR prediction.

C. Impact of Video Content on Consumer Behavior

The semantic features of video advertisements have been widely studied in psychology and marketing fields for their influence on consumer behavior. Zhang et al. [20] employed eye-tracking techniques and found that attention to products and actors in a video ad positively correlates with ad recall and purchase intent, whereas attention to brand logos shows a negative correlation. Wu et al. [21] measured electrical activity in the prefrontal and temporal lobes and reported that manipulating message content and music tempo in video ads affected olfactory and gustatory imagery, which in turn correlated with consumer responses. Based on Gross's emotion regulation model [22], Teixeira et al. [23] used infrared eye-tracking and emotion detection systems to show that joy and surprise are strongly associated with user attention and retention in video advertisements.

These findings highlight the multifaceted ways in which semantic features can shape consumer cognition and behavior. While prior computational approaches to CTR prediction have primarily relied on low-level visual features, these studies suggest that higher-order semantic content plays a critical role in user engagement [20], [21], [23]. Building on this perspective, the present study focuses on extracting semantically meaningful features from video content and leveraging them to improve CTR prediction.

III. Proposed Method

A. Extraction of Semantic Features

As discussed in Section II-C, prior studies have demonstrated that the semantic features of video advertisements can influence consumer responses. However, these findings are often based on specific datasets and controlled settings, making it unclear whether similar insights apply to general-purpose datasets. To address this limitation, we leverage the in-context learning capabilities of foundation models, which enable adaptation to new tasks using only a few labeled examples [24]. Due to the token limitations of the model, we construct a prompt using 20 samples from the training set: 10 with the highest CTR and 10 with the lowest CTR. Each sample consists of 16 video frames arranged in a 4×4 grid, providing a visual summary of the advertisement's temporal flow. To extract a set of frames from each video, we divided each video into 16 equal-duration segments and extracted one

TABLE I
PROMPT FOR IDENTIFYING SEMANTIC FEATURES RELEVANT TO CTR (TRANSLATED FROM JAPANESE).

These images are composed of 16 sequential frames extracted from each video advertisement. They are arranged in a 4×4 grid, starting from the top-left (first frame) and proceeding rightward and downward to the bottom-right (sixteenth frame). The first 10 samples correspond to advertisements with high CTR values, and the latter 10 to those with low CTR values. CTR values for high-CTR ads: 0.XX, 0.XX, 0.XX, ... CTR values for low-CTR ads: 0.XX, 0.XX, 0.XX, ... Based on these examples, please list several semantic features that appear to influence CTR values. {Image 1} {Image 2} ... {Image 20}

frame from the beginning of each segment. This prompt is input to the model at once, and the model returns a single natural language response describing generalizable semantic features across the dataset.

The complete prompt used for in-context learning is shown in Table I. Since the ads are originally in Japanese, the actual prompt was written in Japanese to ensure alignment with the linguistic context of the dataset. Table I presents an English translation of the original prompt. By analyzing the model's response, we automatically identify dataset-specific semantic features that are likely to influence CTR, providing a foundation for constructing generalizable evaluation criteria.

B. CTR Prediction Model Incorporating Semantic Features

We propose a CTR prediction model that integrates semantic features extracted from generated natural language descriptions of video advertisements. As a baseline, we adopt the model proposed by Ikeda et al. [5], which performs regression-based CTR prediction using three modalities: text (ad catchphrase), metadata, and video frames. In contrast, our model incorporates semantics as a fourth modality, as shown in Fig. 1.

For text and semantics, we use a pre-trained Japanese BERT encoder (bert-base-japanese-v3 [25]) to obtain embeddings. For metadata, categorical variables are first processed via one-hot encoding, while numerical variables are passed as-is; these two are then fed separately into layers and concatenated to form a unified embedding. Video inputs are sampled into 16 frames, resized, and fed into a Vision Transformer (ViT) [26], from which we extract the final hidden layer outputs. Each modality-specific embedding is passed through a Fully Connected (FC) layer, a Batch Normalization (BN) layer, and a ReLU layer. In addition, embeddings for video frames and semantic features are further processed by an attention layer. The output of each branch is a 256-dimensional feature vector.

These four modality vectors are integrated using an attention layer, which computes attention weights via an FC layer, a LeakyReLU layer, another FC layer, and a Softmax layer. The output vector is passed through a final prediction block consisting of an FC layer, a BN layer, a ReLU layer, a Dropout layer ($p = 0.5$), and another FC layer to output the predicted

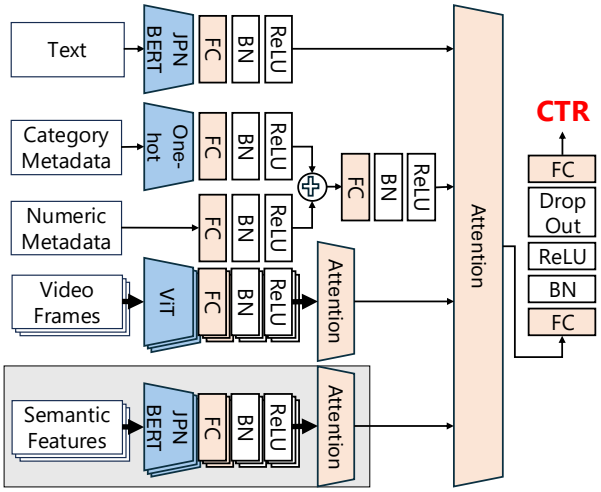


Fig. 1. Model architecture incorporating semantic feature as an additional modality.

CTR $\hat{y}_i$. Let x_i^{text}, x_i^{meta}, x_i^{video}, and x_i^{sem} denote the input vectors for the i-th ad corresponding to the text, metadata, video frames, and semantic features, respectively. The final output is computed as:

$$\hat{y}_i = f_{\boldsymbol{\theta}}(x_i^{\text{text}}, x_i^{\text{meta}}, x_i^{\text{video}}, x_i^{\text{sem}}), \tag{1}$$

where $f_{\boldsymbol{\theta}}$ is a regression model parameterized by $\boldsymbol{\theta}$, which is optimized to minimize the MSE between $\hat{y}_i$ and the ground-truth CTR.

IV. EXPERIMENTS

A. Experimental Setup

We used a TikTok video advertisement dataset owned by Septeni Japan, Inc. The dataset consists of 23,040 training samples and 1,957 evaluation samples, each comprising an MP4-format video file, a textual ad catchphrase, and associated metadata. This dataset includes only impression and click counts per advertisement, and does not contain any personally identifiable information. The training set includes ads published between February 1, 2023, and January 31, 2024, while the evaluation set covers those published between February 1 and February 29, 2024. This temporal split was designed to simulate the task of predicting future ad performance. The advertisements span a wide range of genres, including finance, healthcare, and shopping apps. In terms of video characteristics, the durations range from 5.00 to 60.06 seconds, with an average of 22.58 seconds. Regarding frame rates, the most common setting was 30 frames per second (fps), which accounted for approximately 90% of the videos. A smaller number of videos used 24 fps, 60 fps, or other rates.

We carefully preprocessed the dataset prior to training. First, we removed an entry associated with an account ID that only appeared in the evaluation set, which would interfere with one-hot encoding. Next, we excluded samples with zero clicks, as their CTR is always zero regardless of impression count,

skewing the data distribution. We also removed entries with fewer than 500 impressions, since minor click fluctuations could cause unstable CTR values and negatively affect training. After these steps, the evaluation set was split into a validation set and a final evaluation set.

As a result, we obtained 19,503 samples for training, 754 for validation, and 754 for evaluation. The resulting data split is skewed toward training, with 92.8% of samples allocated to training, and 3.6% each to validation and evaluation. This imbalance arises from the dataset's predefined structure and our need to allocate part of the evaluation set for validation while maintaining the integrity of future data prediction. In this study, we prioritized the goal of predicting unseen video ad performance and thus opted to use part of the evaluation set as validation data. Finally, since the distribution of raw CTR values is heavily skewed toward zero, we applied the log transformation defined in (2) and used the result as the regression target.

$$\log_{10}(100 \times \mathrm{CTR} + 1). \tag{2}$$

We employed the Adam optimizer and used MSE as the loss function. The number of training epochs was set to 50, and the learning rate was decayed by a factor of 0.95 after each epoch. To determine hyperparameters, we used the Optuna library [27], which performs Bayesian optimization. The search space included the initial learning rate, sampled on a logarithmic scale between 10^{-3} and 10^{-5}, and the batch size, selected from $\{64, 128, 256\}$. Models were trained using NVIDIA A100 GPUs, taking approximately 10 hours in total.

To incorporate semantic features into our model, we first extracted a set of features that characterize CTR using the in-context learning prompt (Table I) described in Section III-A. Table II lists six semantic features that appeared in the outputs when the same prompt was executed five times, and which occurred in at least two of those outputs. In contrast to a naive prompt that simply asks in general terms—e.g., "Please list several features in the content of video advertisements that are likely to influence CTR"—the in-context prompt enabled the identification of two additional categories: "Message Clarity" and "Incentive." This result suggests that prompts with contextual examples can help reveal latent factors strongly associated with CTR in this dataset. Based on these six features, we constructed a systematic prompt (Table III) to extract semantic features from each video in the dataset for both training and evaluation sets. Using the descriptions generated by this prompt as additional model input, we trained and compared 64 models in parallel across all 2^6 combinations of the six features.

B. Quantitative Results

Due to confidentiality agreements with Septeni Japan, Inc., we have to avoid to disclose the absolute values of CTR prediction errors. Therefore, all quantitative results are reported and discussed in terms of relative metrics.

Table IV shows the relative improvement in MSE for all 64 model variants, compared to the baseline model without semantic features. Each row represents a different combination of semantic features, and the improvement is computed as the percentage reduction in MSE from the baseline. Paired t-test p-values versus the baseline are also reported. Among the 63 semantic models, 62 achieved better performance than the baseline. The best-performing model showed a 14.02% reduction in MSE, demonstrating the effectiveness of incorporating semantic features in CTR prediction.

Table V presents the average MSE improvement for each added semantic feature. For each feature, we averaged the MSE differences across 32 model pairs that differ only by the presence or absence of that feature—for example, comparing the model using features 2–6 with the model using features 1–6 is a pair to evaluate feature 1. Additionally, paired t-test p-values are reported. As shown in Table V, "Message Clarity" and "Incentive" yielded the most substantial improvements,

TABLE II
SEMANTIC FEATURES CHARACTERIZING CTR.

Semantic Feature	Description
1. Visual Attractiveness	Presence of vivid colors, striking design, or eye-catching visuals.
2. Message Clarity	Whether the message is concise and clearly communicated.
3. Brand Recognition	Use of well-known brands or characters.
4. Incentive	Presence of specific incentives or benefits.
5. Storytelling	Whether the ad has a narrative or emotional arc.
6. Call-to-Action (CTA)	Features that encourage user interaction, such as buttons or links.

TABLE III
PROMPT FOR EXTRACTING SEMANTIC FEATURES FROM THE DATASET (TRANSLATED FROM JAPANESE).

This image combines 16 frames extracted from a video advertisement in chronological order, arranged in a 4×4 grid. The top-left is the first frame, and it continues row-wise to the bottom-right as the 16th.
Please evaluate the following six aspects and, if applicable, point out their shortcomings. Replace the placeholders in parentheses as instructed.
#1 (Does the ad use vivid colors or visuals to attract attention?)
...
{Image}

TABLE IV
MSE IMPROVEMENTS ACROSS DIFFERENT COMBINATIONS OF SEMANTIC FEATURES.

Rank	Features Used	MSE Improvement [%] ↑	P-value
1	2	14.02	0.001
2	2, 3, 4, 5, 6	13.12	0.007
...	...	...	...
63	None (Baseline)	0.00	N/A
64	2, 3	−0.08	0.992

TABLE V
AVERAGE MSE IMPROVEMENTS OVER PRESENCE OF EACH FEATURE.

Added Feature	Avg. MSE Improvement [%] ↑	P-value
4. Incentive	2.10	0.021
2. Message Clarity	1.51	0.075
6. Call-to-Action (CTA)	1.22	0.180
5. Storytelling	0.20	0.801
1. Visual Attractiveness	−0.64	0.395
3. Brand Recognition	−1.15	0.022

TABLE VI
PREDICTION ERRORS BEFORE AND AFTER ADDING "MESSAGE CLARITY".

Ad Example	Without Feature	With Feature
Fig. 2	-0.6603	**-0.2169**
Fig. 3	**-0.0411**	-0.3725

supporting our hypothesis that in-context learning enables the extraction of dataset-specific semantic features. In contrast, the addition of "Visual Attractiveness" and "Brand Recognition" worsened prediction accuracy, possibly because their descriptions tended to use generic phrases like "colorful" or "famous brand", limited the ability to identify meaningful predictors of CTR.

C. Case Study

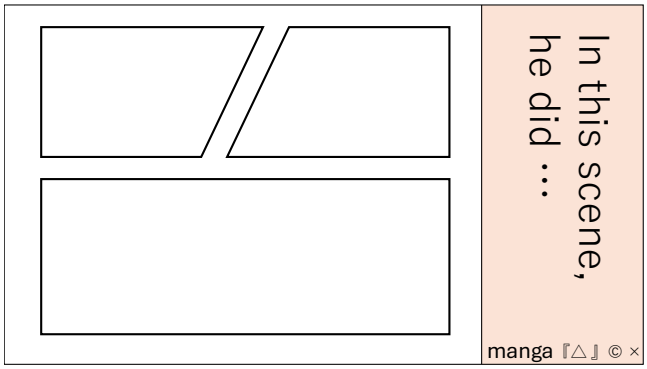
To examine how semantic features affect model predictions, we conducted a case study focusing on the "Message Clarity" feature. We selected two representative examples in which the addition of this feature either significantly improved or worsened prediction accuracy, as illustrated in Figs. 2 and 3. Note that, due to copyright issues, these figures present simplified illustrations rather than the actual video frames. Table VI summarizes the relative prediction errors for each case, comparing results from the baseline model (without the feature) and the updated model (with the feature).

In Fig. 2, the advertisement conveys a simple and repetitive message such as "This is the definitive horse racing app!", making the intent of the ad easy to understand. The generated description correctly captured this high level of clarity. Accordingly, the model incorporating the "Message Clarity" feature predicted a CTR value closer to the ground truth than the baseline model.

In contrast, the ad in Fig. 3 is for a manga app, but shows fragmented text sequentially, and it is only at the end that the app's purpose becomes clear. Generated description reflected this low clarity, and the model prediction incorporating the "Message Clarity" feature appropriately lowered the predicted CTR. However, since the baseline model's prediction was already lower than the ground truth, the addition of the semantic feature led to a larger error. These examples suggest that incorporating semantic features such as "Message Clarity" helps the model make more contextually grounded predictions.

GPT-4o Description – Message Clarity:
The message is clear, with the phrase "This is the definitive horse racing app!" repeated throughout, making it easy to understand. The information is concise, and the message comes across as sufficiently strong.

Fig. 2. Example where prediction accuracy improved with the addition of the "Message Clarity" feature (translated from Japanese).



GPT-4o Description – Message Clarity:
The advertisement message is short and direct, but it might be even better if it were designed to make the information easier to understand at a glance.

Fig. 3. Example where prediction accuracy worsened with the addition of the "Message Clarity" feature (translated from Japanese).

V. CONCLUSION

This study proposed a method for enhancing CTR prediction of video advertisements by incorporating semantic features extracted via a foundation model. Unlike conventional approaches that primarily rely on metadata or user logs, our method evaluates the meaning and structure of the ad content itself. Using GPT-4o and in-context learning, we extracted six interpretable semantic features such as "Message Clarity" and "Incentive." Experimental results showed that incorporating these features led to improved prediction accuracy in most cases, with the best-performing model achieving a 14.02% reduction in MSE. Additionally, we demonstrated that in-context learning prompts enable the automatic discovery of dataset-specific semantic features that are strongly associated with CTR. A case study illustrated how the "Message Clarity" feature contributed to both improving and worsening prediction performance.

Overall, this work presents a framework for bridging the gap between qualitative ad content and quantitative prediction, offering a foundation for more explainable and content-aware advertisement optimization. One current limitation is that our method treats video content as static, without accounting for temporal variations in semantic features. To address this, we are currently working on modeling the sequential evolution of semantic features across video timelines. Beyond CTR prediction, our future work aims to generate improved narrative structures that enhance CTR, and ultimately to synthesize video advertisements based on those narratives.

References

[1] Technavio, *Digital Video Advertising Market Analysis North America, APAC, Europe, South America, Middle East and Africa – US, China, UK, Germany, Japan – Size and Forecast 2024–2028*, https://www.technavio.com/report/digital-video-ad-market-analysis, Accessed: 2025-06-16, 2024.

[2] Nielsen, *Nielsen Study Reveals Majority of Consumers Actively Avoid Ads Across Podcasts, Streaming, and Live TV Platforms*, https://www.nielsen.com/news-center/2023/nielsen-study-reveals-majority-of-consumers-actively-avoid-ads-across-podcasts-streaming-and-live-tv-platforms/, Accessed: 2025-06-16, 2023.

[3] Y. Yang and P. Zhai, "Click-Through Rate Prediction in Online Advertising: A Literature Review," *Information Processing and Management*, vol. 59, no. 2, p. 102 853, 2022.

[4] L. He, H. Chen, D. Wang, S. Jameel, P. Yu, and G. Xu, "Click-Through Rate Prediction with Multi-Modal Hypergraphs," in *CIKM*, 2021, pp. 690–699.

[5] J. Ikeda, H. Seshime, X. Wang, and T. Yamasaki, "Predicting Online Video Advertising Effects with Multimodal Deep Learning," in *ICPR*, 2021, pp. 2995–3002.

[6] J. Lin, X. Dai, Y. Xi, *et al.*, "How Can Recommender Systems Benefit from Large Language Models: A Survey," *ACM Transactions on Information Systems*, vol. 43, no. 2, 2024.

[7] M. Richardson, E. Dominowska, and R. Ragno, "Predicting Clicks: Estimating the Click-Through Rate for New Ads," in *The Web Conference*, 2007, pp. 521–530.

[8] H. B. McMahan, G. Holt, D. Sculley, *et al.*, "Ad Click Prediction: A View from the Trenches," in *SIGKDD*, 2013, pp. 1222–1230.

[9] G. Zhou, X. Zhu, C. Song, *et al.*, "Deep Interest Network for Click-Through Rate Prediction," in *SIGKDD*, 2018, pp. 1059–1068.

[10] A. Vaswani, N. Shazeer, N. Parmar, *et al.*, "Attention Is All You Need," in *NeurIPS*, 2017, pp. 6000–6010.

[11] T. Brown, B. Mann, N. Ryder, *et al.*, "Language Models Are Few-Shot Learners," in *NeurIPS*, 2020, pp. 1877–1901.

[12] F. Petroni, T. Rocktäschel, P. Lewis, *et al.*, "Language Models as Knowledge Bases?" In *EMNLP*, 2019, pp. 2463–2473.

[13] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding," in *NAACL*, 2019, pp. 4171–4186.

[14] B. Geng, Z. Huan, X. Zhang, *et al.*, "Breaking the Length Barrier: LLM-Enhanced CTR Prediction in Long Textual User Behaviors," in *SIGIR*, 2024, pp. 2311–2315.

[15] J. Lin, B. Chen, H. Wang, *et al.*, "ClickPrompt: CTR Models Are Strong Prompt Generators for Adapting Language Models to CTR Prediction," in *The Web Conference*, 2024, pp. 3319–3330.

[16] D. Ko, S. Lee, and G. Kim, "Can Language Models Laugh at YouTube Short-form Videos?" In *EMNLP*, 2023, pp. 2897–2916.

[17] K. Ataallah, X. Shen, E. Abdelrahman, *et al.*, "MiniGPT4-Video: Advancing Multimodal LLMs for Video Understanding with Interleaved Visual-Textual Tokens," *arXiv preprint arXiv:2404.03413*, 2024.

[18] G. DeepMind, *Gemini assistant*, https://gemini.google/assistant/, Accessed: 2025-06-18, 2023.

[19] OpenAI, *Hello gpt-4o*, https://openai.com/index/hello-gpt-4o/, Accessed: 2025-06-18, 2024.

[20] X. Zhang and S.-M. Yuan, "An Eye Tracking Analysis for Video Advertising: Relationship Between Advertisement Elements and Effectiveness," *IEEE Access*, vol. 6, pp. 10 699–10 707, 2018.

[21] Y.-L. Wu and P.-C. Chen, "Neurophysiology of Sensory Imagery: An Effort to Improve Online Advertising Effectiveness Through Science Laboratory Experimentation," *Information & Management*, vol. 61, no. 4, p. 103 708, 2022.

[22] J. J. Gross, "Emotion Regulation: Conceptual and Empirical Foundations," *Handbook of Emotion Regulation*, vol. 2, pp. 3–20, 2014.

[23] T. Teixeira, M. Wedel, and R. Pieters, "Emotion-Induced Engagement in Internet Video Advertisements," *Journal of Marketing Research*, vol. 49, no. 2, pp. 144–159, 2012.

[24] P. Liu, W. Yuan, J. Fu, Z. Jiang, H. Hayashi, and G. Neubig, "Pre-Train, Prompt, and Predict: A Systematic Survey of Prompting Methods in Natural Language Processing," *ACM Computing Surveys*, vol. 55, no. 9, pp. 1–35, 2023.

[25] Tohoku NLP Group, *Tohoku-nlp/bert-base-japanese-v3*, https://huggingface.co/cl-tohoku/bert-base-japanese-v3, Accessed Mar 9, 2025, 2021.

[26] A. Dosovitskiy, L. Beyer, A. Kolesnikov, *et al.*, "An Image Is Worth 16×16 Words: Transformers for Image Recognition at Scale," in *ICLR*, 2021.

[27] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A Next-Generation Hyperparameter Optimization Framework," in *SIGKDD*, 2019, pp. 2623–2631.