

Web Appendix to Music Marketing with Generative AI

A Related Literature	iii
B Marketing Asset Generation Engine (MAGE)	iv
B.1 Minimal Input Requirements	iv
B.2 Standardized Design Templates	iv
B.3 Random Parameter Sampling	x
C Overview of Studies	xii
D Study 1 Extension: Balanced 1 AI vs. 1 Human A/B Test	xiii
E Audience Composition and Divergent Delivery	xv
E.1 Audience Composition Across Conditions	xv
E.2 Divergent Delivery	xvi
F Model-Free Evidence	xix
G Meta Platform Split: Instagram vs. Facebook	xx
H Waves	xxii
I Expert Selection	xxiii
I.1 Selection Studies I and II: Marketing Professionals and Laypeople	xxiii
I.2 Selection Study III: Industry Experts	xxv
I.3 Selection Study IV: Rank Ordering	xxvi
I.4 Selection Study V: Selection by Engagement	xxviii
J Segment-Specific Content Creation	xxxi
J.1 Training an Evaluative AI Model for AI Video Selection	xxxi
J.2 Applying the Evaluative AI Model for AI Video Selection	xxxi
K Cost Per Mille (CPM)	xxxii
L Pooled Results Across Studies 1-3	xxxiii
L.1 Ad-level Results	xxxiii
L.2 Pooled Results Across Studies 1-3 with Study Interactions	xxxv
L.3 Winsorization	xxxvi

M Lab Studies: Replication and Heterogeneous Effects	xxxvii
N Heterogeneous Effects	xl
N.1 Heterogeneous Effects of AI Video Ads Across Audiences	xl
N.2 Heterogeneous Effects of AI Video Ads Controlling for Release Year	xli
N.3 Age Gradient for CTR Effect of AI Video Ads	xliii
N.4 Analysis of Visual Features and Audience Response	xliii
O Robustness Analyses	xlvi
O.1 Reach	xlvi
O.2 Aggregated Demographics	xlix
O.3 OLS	l
O.4 Winsorization	li
P Visual Content Similarity	liv
Q Quantifying the Business Impact of MAGE	lvii
R Landing Page Engagement	lix
S Genre	lx

A Related Literature

Table A1 compares prior work on AI-generated advertising across input and output modalities, advertising platforms, number of campaigns, use of performance data for generation or selection, and analysis of demographic heterogeneity.

Source	Input modality	Output modality	Ad platforms	Number of campaigns	Performance informed ^a	Demographic heterogeneity ^b
This paper	Music, text, image	Video	TikTok, Meta (Instagram, Facebook)	75	Yes (evaluative AI)	Yes (age, gender)
Jiang et al. (2025)	Text	Text	Meta (Facebook)	1	Yes (data curation)	-
Dubé and Xu (2026)	Text	Text	Email	3	-	-
Reisenbichler et al. (2022)	Text	Text	Google	2	Yes (data curation)	-
Reisenbichler et al. (2025)	Text	Text	Google	2	Yes (data curation)	-
Fang et al. (2025)	Text	Text	Google, App	Undisclosed	-	-
Chen et al. (2025)	Text	Text	Douyin Search	Undisclosed	Yes (evaluative AI)	-
Chen and Chan (2024)	Text	Text	Meta (Instagram, Facebook)	355	-	-
Angelopoulos et al. (2024)	Text	Text	Undisclosed	36	Yes (data curation)	-
Hunt and Kerr (2025)	Text	Text	Meta ^c	12	Yes (data curation)	-
Exner et al. (2025)	Text	Text, image	Taboola	351 ^d	-	-
Hartmann et al. (2025)	Text	Image	Meta (Facebook)	2	-	-
To et al. (2025)	Text	Image	Email	4	-	-
Heitmann et al. (2025)	Text	Image	Meta ^c , Taboola	3	Yes (data curation)	-
Lee et al. (2025)	Text, image	Image	Google	5	-	-
Daviet and Nishimura (2026)	Image	Image	Meta (Instagram)	9	Yes (evaluative AI)	-
Efthymiou et al. (2024)	Text	Voice	YouTube	4	-	-
Kapoor and Kumar (2025)	Text	Video (avatar), image	WhatsApp	1	-	Yes (gender)

Table A1: Overview of academic literature on AI-generated advertising

Notes: All references listed use real-world field data and were identified through a structured keyword search of recent academic work on AI-generated advertising, using major electronic databases as well as references from selected sources (Exner et al., 2025; Jiang et al., 2025).

^a Yes, if real-world performance data (e.g., CTR) inform the generation or selection of AI-generated ads. ‘Evaluative AI’ denotes a model trained on past performance to predict the performance of candidate ads, enabling automated, data-driven selection. ‘Data curation’ denotes the use of a performance metric to filter training examples, steering a generative AI model’s training process towards real-world performance.

^b Yes, if demographic variables serve as moderating variables in a dedicated analysis of heterogeneous effects, and not just as controls

^c Undisclosed if Instagram and/or Facebook

^d Observational (“sibling ads”)

B Marketing Asset Generation Engine (MAGE)

B.1 Minimal Input Requirements

Key content serves as a fixture for the campaign, directly impacts all other assets, and is intrinsically associated with the product (e.g., the music track, album cover, music video, or artist photos). Derived content is generated by editing, adapting, or expanding key content, resulting in artwork animations, alternative sound treatments, or lyric videos. MAGE takes key content as input and produces derived content as output.

For the visual input, MAGE relies on the high-quality, typically quadratic packshot $\mathcal{I}$, denoted by the (height $\times$ width $\times$ color) three-dimensional tensor of pixels in which color $\mathcal{C} := (C_R, C_G, C_B) \in [0, 255]^3 \subset \mathbb{R}^3$ in the RGB color space. For example, typical inputs include key content such as a 3000×3000 px high-resolution .tiff file for the packshot (~ 150 - 250 MB). The music metadata $\mathcal{T} := \{\mathcal{T}^{track}, \mathcal{T}^{time}\}$ reflect a string that includes the name of the track $\mathcal{T}^{track}$ and a timestamp $\mathcal{T}^{time}$ to mark the beginning of an audio thumbnail, which constitutes the most representative section of the track. The timestamp is in the format 00:00:00.000 (hh:mm:ss.msmsms). Finally, the audio input $\mathcal{A}$ a high-resolution audio file of the track, encoded in an uncompressed .wav format (~ 50 MB). MAGE then outputs files according to platform-specific requirements, such as .mp4 video files with 1080×1920 px resolution.

B.2 Standardized Design Templates

The implementation of MAGE is based on a channel definition and a design template definition. The channel definition ensures fulfillment of technical platform requirements, such as resolution, video codec, or file formats. The design template definition encompasses the creative components of the generated asset, including background elements, text displays, and audio implementation. On a single channel, multiple different contents can be generated, and a single content can be deployed across different platforms.

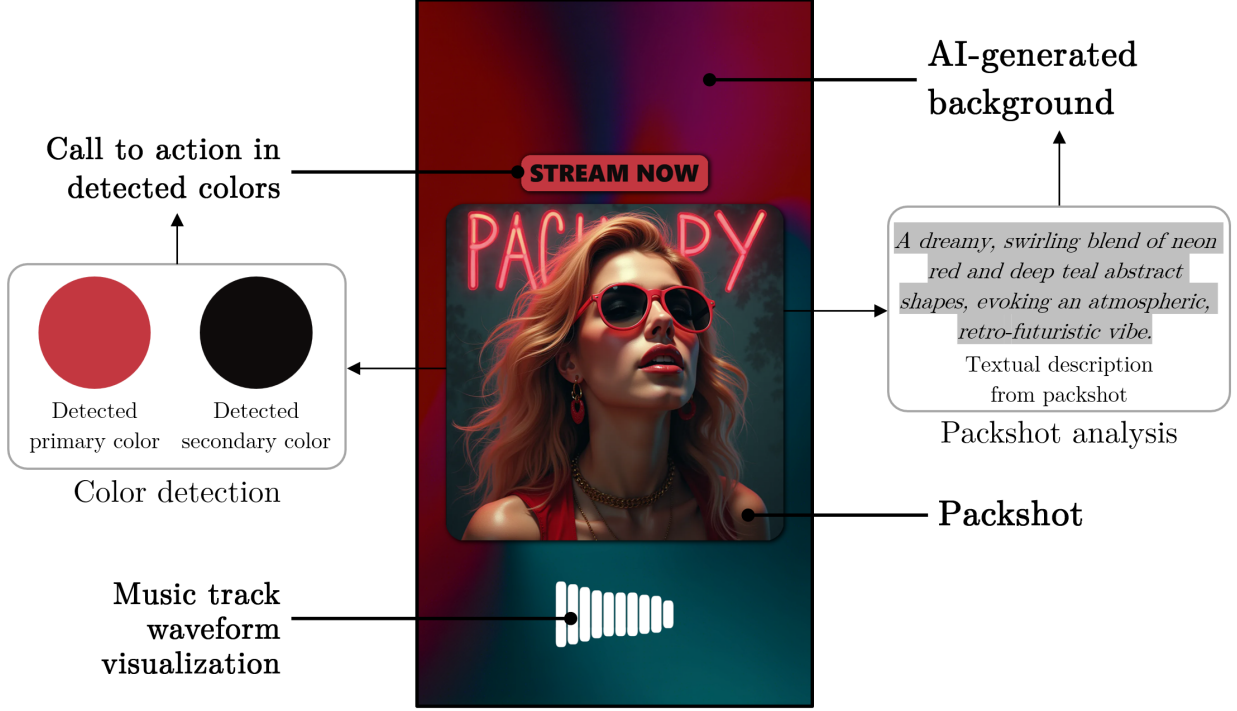


Figure A1: Components of an example AI video ad

Figure A1 depicts the elements of an exemplary AI video ad. The AI-generated background is derived from a prompt informed by a textual description of the packshot, obtained through the OpenAI API for GPT-4o (OpenAI et al., 2023). For background generation, we implement state-of-the-art generative AI models for image and video synthesis, specifically Stable Diffusion Ultra (Esser et al., 2024) and Stable Video Diffusion (Blattmann et al., 2023). The prompt used to extract the packshot description is “One sentence image prompt for a slightly abstract video background for this packshot with the same color palette and without any names of song, band, or album”.

The animation engine animates the packshot in multiple ways for visual appeal and adds text, i.e., “Stream now”. The music engine extracts timestamped audio clips $\tilde{\mathcal{A}}$ from $\mathcal{A}$, which MAGE synchronizes with animated elements (e.g., waveform visualization). The system generates multilayered videos with foreground, background, and audio, producing final assets $\mathcal{V}_i^{Ani}$ through frame-by-frame animation with video-music synchronization.

The call to action (CTA) appears as text within a text box, with varied position (above or

below the packshot), color and font size. All assets show the text “Stream now”. Colors are determined by a custom-developed machine learning algorithm that analyzes the packshot to identify primary and secondary colors. First, MAGE identifies matching colors on the basis of the visual input $\mathcal{I}$, such that it ensures visual congruence between the packshot and other elements (Henderson et al., 2019). Therefore, we apply k-means clustering to all pixels of the image in the pixel space to obtain relevant colors (Celebi, 2011). For the selections of the primary and secondary colors, $\mathcal{C}_1$ and $\mathcal{C}_2$, we identify the first color with intensity values (i.e., difference between the maximum and minimum R, G, and B values) above and below a threshold, resulting in one vivid and one neutral color. These detected colors appear in the CTA text, backgrounds, and intro/outro shots (see Figure A2 and 1).

In addition, with the assistance of a multimodal large language model (LLM) $\mathcal{M}_{LLM}$, a predefined text prompt $\mathcal{T}_{description-prompt}$, and an image $\mathcal{I}$, we obtain a textual description of the packshot $\mathcal{I}$, which takes the form of the text output $\mathcal{T}_{description} = \mathcal{M}_{LLM}(\mathcal{T}_{description-prompt}, \mathcal{I})$.

For example, an image $\mathcal{I}^{Gen}$ generated with the $\mathcal{M}_{Diffuser}^{T2I}$ visually represents the prompt $\mathcal{T}_{prompt}$: $\mathcal{I}^{Gen} = \mathcal{M}_{Diffuser}^{T2I}(\mathcal{T}_{prompt})$. For video generation, we also can use a generated image $\mathcal{I}^{Gen}$ as input for image-to-video generation, such that $\mathcal{M}_{Diffuser}^{I2V}$ would result in a video $\mathcal{V}^{Gen} = \{\mathcal{I}_{\mathcal{V}^{Gen},1}, \dots, \mathcal{I}_{\mathcal{V}^{Gen},N_f}\}$, reflecting a sequence of N_f frames: $\mathcal{V}^{Gen} = \mathcal{M}_{Diffuser}^{I2V}(\mathcal{I}^{Gen})$. The overall generative model $\mathcal{M}_{Diffuser}$ therefore can be described as:

$$\mathcal{V}^{Gen} = \mathcal{M}_{Diffuser}(\mathcal{T}_{prompt}) := \mathcal{M}_{Diffuser}^{I2V}(\mathcal{M}_{Diffuser}^{T2I}(\mathcal{T}_{prompt})). \quad (\text{A1})$$

The packshot itself, which may incorporate resizing, rounded edges, and shadow effects, is integrated into the center of the frame. Position correction is applied when necessary to balance the weight of the call to action and waveform, e.g., when call to action is placed below the packshot. A music track waveform visualization can be displayed at the bottom of the frame. Figure A2 presents a range of example AI assets and a human asset, all showcased by multiple frames. The AI assets illustrate the creative variability that MAGE facilitates

while preserving each artist’s distinct visual language.

The parameters Θ_i and the inputs $\mathcal{D}$ provide a complete representation of the generated output video $\mathcal{V}_i$, with information that define its characteristics.

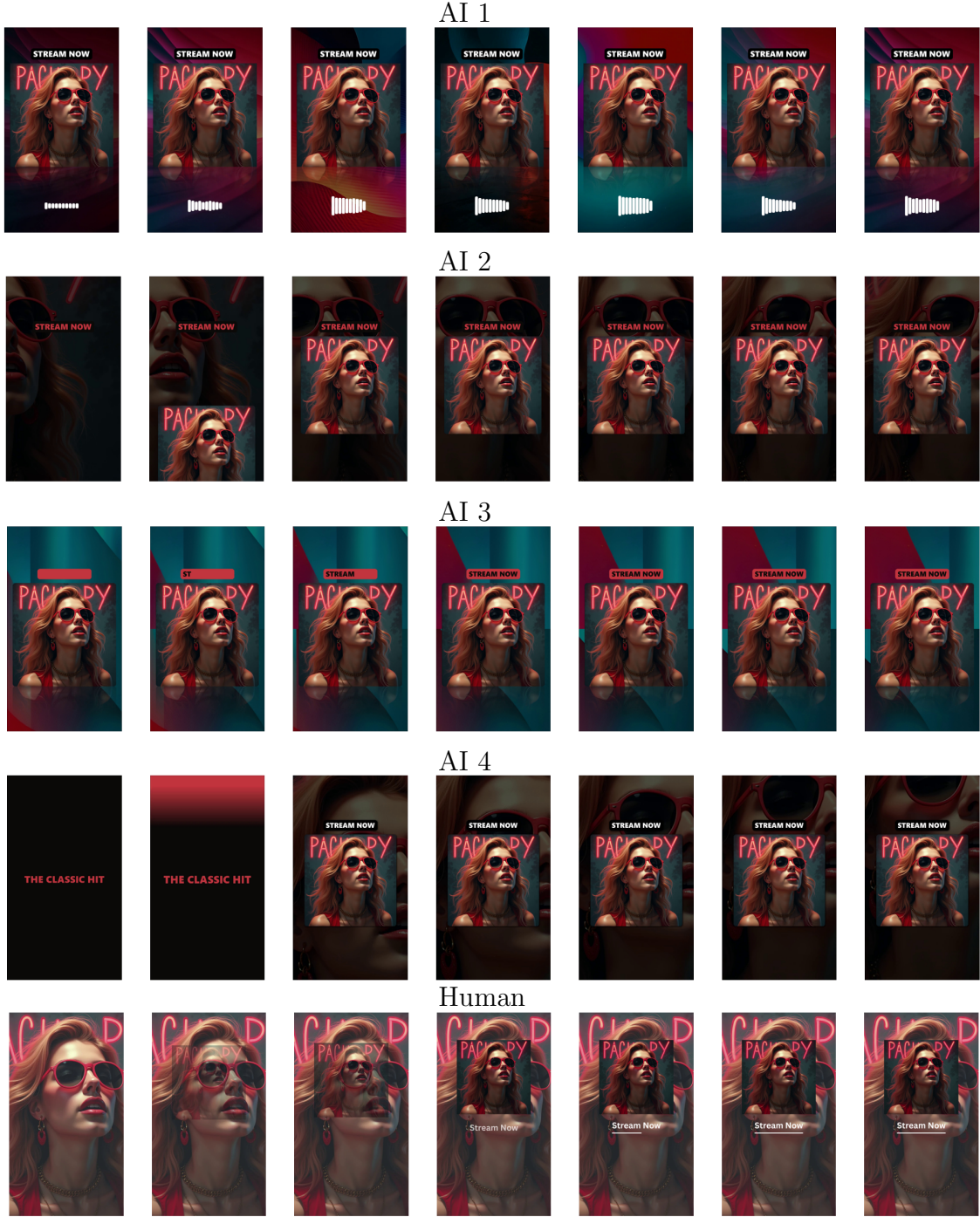


Figure A2: Examples of video ads

Tables A2 and A3 summarize all video elements of MAGE, illustrating the full range of

creative components available for asset generation.

ID	Video category	Variant	Description
1	Intro/Outro	Intro	Short, separate intro shot based on selected colors, showing the text <i>THE CLASSIC HIT</i> .
2	Intro/Outro	Outro	Short, separate outro shot based on selected colors, showing the text <i>THE CLASSIC HIT</i> $\{track-name\}$ and afterwards <i>Listen now</i> .
3	Content	Packshot position	The packshot can be shown in the middle or with an offset to either top or bottom.
4	Content	Packshot size	The packshot can be shown in different sizes.
5	Content	Packshot edges	The edges of the packshot can be rounded with different radii.
6	Content	Packshot shadow	The packshot can be shown with or without shadow.
7	Content	No animation	Video foreground is always based on a scaled version of the packshot (which is given as input). In this case, no additional animation is applied.
7	Content	Boom animation	The packshot is moved according to the beats detected in the audio, i.e., the moment a beat is detected, a pulsing effect is applied for some frames.
7	Content	Fly in animation	The video starts without the packshot, which then moves inside the frame from the bottom.
7	Content	Product animation	In this case, the video starts with the packshot only, but then a vinyl is animated rotating and partially moving out of the packshot.
7	Content	Reflection animation	This animation consists out of a partial reflection of the packshot to the bottom.
7	Content	Shear animation	An animation that applies a shear effect to the packshot, resulting in a slight 3D movement.
7	Content	Shiny animation	A shiny effect is swepted diagonally across the packshot.
7	Content	Zoom animation	The packshot is initially zoomed-in and afterwards zoomed-out again.
8	Content	Waveform	There can be a waveform visualization (visualizing the audio) at the bottom of the frame.

Table A2: Descriptions of video parameters (1/2)

ID	Video category	Variant	Description
9	Background	Color	Video background can be either solid (one color) or as radial or angled gradient (animated, two colors).
9	Background	Packshot	Video background based on the packshot, which is scaled and shown with lower visibility and animated with rotation or movement.
9	Background	AI image	Video background based on AI images, either one image (GenAI Image), two images (Split Images) at the same time (top and bottom), or four images (GenAI Images) after one another. Animated backgrounds with rotation or movement.
9	Background	AI video	Video background based on AI videos, either four videos (GenAI Videos) after one another or 2× the amount, i.e., 8, videos (GenAI 2× Videos).
10	Background	Animation	Static image backgrounds can be animated with movement and rotation.
11	Background	Prompt	Generated backgrounds are created using a modified text prompt.
12	Background	Motion Bucket ID	The pace of generated videos is controlled by the motion bucket.
13	Background	Brightness	The background can be in a lighter or darker form
14	Text	Text position	Text can either be above or below the packshot.
15	Text	Font size	The font size of can be small, middle or large.
16	Text	Color	Text and its background can be in different (e.g., packshot-dependent) colors.
17	Text	Animation	Text can either be static (present all the time) or animated letter by letter (starting without the text and only the textbox can be seen).
18	Audio	Loudness	Audio can be in original or in increased loudness.
19	Audio	Speed	Can be slow (90%), original, or fast (105% or 110%).

Table A3: Descriptions of video parameters (2/2)

B.3 Random Parameter Sampling

The creative elements differ across the final assets, reflecting their sampled parameter vectors Θ_i , but the assets remain consistent with the visual language of the artist:

$$\mathcal{V}_i = \mathcal{B}(\Theta_i, \mathcal{D}) = \mathcal{B}(\Theta_i, (\mathcal{T}, \mathcal{I}, \mathcal{A})). \quad (\text{A2})$$

The sampling process is governed by a probability distribution $P(\theta_k = p_{kl}) = w_{kl}$, $l = 1, 2, \dots, M_k$, where w_{kl} , with $\sum_{l=1}^{M_k} w_{kl} = 1$ and $w_{kl} \geq 0$, are the sampling weights for the l -th value in $\mathcal{P}_k$, which has M_k entries. Then it is possible to obtain the full set of parameters Θ by sampling $\theta_k \sim P(\mathcal{P}_k, \mathbf{w}_k)$, where $\mathbf{w}_k = \{w_{k1}, w_{k2}, \dots, w_{kM_k}\}$ are the weights associated with the sampling distribution for $\mathcal{P}_k$.

MAGE can create more than 783 million distinct variants through its sampling engine. This expansive creative space results from independently sampling most parameters while imposing minimal restrictions on available options, as illustrated in Figure 3. Note the number of variants could be much higher if more parameter variations would be permitted or multiple generations with the same prompt would be allowed. Table A4 summarizes all parameters that are sampled as part of MAGE, along with their respective sampling probabilities as implemented in Study 3.

ID	Feature	Variations	Probabilities
1	Intro	[Yes, No]	[12.5%, 87.5%]
2	Outro	[Yes, No]	[12.5%, 87.5%]
3	Packshot position	[Top, Middle, Bottom]	Uniform
4	Packshot size	[Small, Medium, Large]	Uniform
5	Packshot edges	Rounded with r= [0, 25, 50]	Uniform
6	Packshot shadow	[Yes, No]	Uniform
7	Animation type	[None, Fly-in, Product, Shear, Reflection, Boom, Shiny, Zoom]	[18.75%, 18.75%, 6.25%, 0%, 18.75%, 0%, 18.75%, 18.75%]
8	Waveform	[Yes, No]	[25%, 75%]
9	Background variant	[Color, Packshot, GenAI Image, Split Images, GenAI Images, GenAI Videos, GenAI 2× Videos]	[0%, 10%, 10%, 10%, 10%, 30%, 30%]
10	Background animation	[Movement (four directions, three velocities), rotation (two directions, two velocities)]	Uniform (Image variants only)
11	Background prompt	Modified for [Patterns (4), Color (4), Light (3), Movement (3)]	Uniform (AI variants only)
12	Motion Bucket ID	[127, 255]	Uniform (AI video variants only)
13	Brightness	[Light, Dark]	[0%, 100%]
14	Text position	[Above, Below] packshot	Uniform
15	Text font size	[Small, Medium, Large]	Uniform
16	Text color	[First color, Second color]	Uniform
17	Text animation	[Yes, No]	Uniform
18	Audio loudness	[Original, Louder]	Uniform
19	Audio speed	[Slow, Original, Fast]	Uniform

Table A4: Parameter sampling variations and probabilities

Notes: Based on expert feedback, some combinations were excluded as part of the design template, e.g., certain background variant and animation combinations.

C Overview of Studies

Table A5 provides an overview of each study’s research question, metrics, and study setup.

Study	Research Question	Main Metrics	Study Setup
Study 1	How effective are AI video ads for music marketing on social media compared to human-created ads by a large music label?	• <i>CTR</i> • <i>ER</i>	Field study (AsPredicted #188218) • Scope: 50 campaigns with 25 music tracks by 22 artists • Conditions: Five video ads (one human, four expert selected AI) per campaign • Exposure: 7.8 million impressions • Budget: GBP 250 per campaign, fixed • Timing: Three waves with 12 campaigns (September 10-12, 2024), 18 campaigns (September 17-19, 2024), and 20 campaigns (October 29-31, 2024)
Study 1 Extension	How do AI video ads perform in a 1 vs. 1 A/B testing setup on TikTok?	• <i>CTR</i> • <i>ER</i>	Field study (AsPredicted #247209) • Scope: Three campaigns with three music tracks by three artists • Conditions: Two video ads (one human, one AI) per campaign • Exposure: .94 million impressions • Budget: GBP 500 per campaign, fixed • Timing: One wave (September 23-25, 2025)
Study 2	Does adding a human in the loop to select AI video ads from a candidate set add value?	• <i>CTR</i> • <i>ER</i>	Field study (AsPredicted #200286) • Scope: 14 campaigns with seven music tracks by four artists • Conditions: Nine video ads (one human, four expert selected AI, four randomly sampled AI) per campaign • Exposure: 2.2 million impressions • Budget: GBP 250 per campaign, dynamic • Timing: One wave (December 3-5, 2024)
Study 3	Can an evaluative AI model predict the performance of segment-specific AI video ads for different target groups before the campaign launch?	• <i>CTR</i> • <i>ER</i>	Field study (AsPredicted #204872) • Scope: Eight campaigns with one music track and eight target audiences • Conditions: Nine video ads (one human, four predicted best AI, four predicted worst AI) per campaign • Exposure: 1.1 million impressions • Budget: GBP 250 per campaign, dynamic • Timing: One wave (December 17-19, 2024)
Lab Study 1	Do the results of Study 1 replicate in a controlled ‘near-exact replication’ in the absence of algorithmic targeting?	• <i>Click intention</i> • <i>Engagement</i>	Lab study (between-subjects, near-exact replication) • 25 AI and 25 human video ads • ~10 participants per condition (492 total)
Lab Study 2	Do heterogeneous effects of AI video ads exist across different audience demographics?	• <i>Click intention</i> • <i>Engagement</i>	Lab study (between-subjects, near-exact replication, AsPredicted #242528) • Six AI and six human video ads • ~100 participants per condition (1,185 total)

Table A5: Summary of studies

Notes: Refer to Table 3 for an overview of selection studies, with details reported in Web Appendix I.

D Study 1 Extension: Balanced 1 AI vs. 1 Human A/B Test

Given that four AI assets compete against a single human asset in each campaign, one might wonder whether the AI video ads benefit through speaking a similar visual language. To address this question empirically, we run a preregistered follow-up study for three music tracks (AsPredicted #247209) comparing AI vs. human ads in a 1 vs. 1 setup on TikTok. To ensure sufficient observations, each asset receives an equal budget of GBP 250, resulting in a total of 937,070 impressions. This setup also enables the use of TikTok’s A/B testing functionality, which is unavailable for campaigns with more than two assets.¹ Table A6 presents the audience composition.

Condition	Impressions	TikTok						
		Gender		Age				
		female	male	18–24	25–34	35–44	45–54	55+
Overall	937,070	55.7%	44.2%	34.2%	20.0%	19.8%	16.3%	9.7%
Human	477,008	55.0%	44.9%	33.7%	19.7%	19.7%	16.7%	10.1%
AI	460,062	56.5%	43.4%	34.7%	20.3%	19.9%	15.9%	9.2%

Table A6: Audience composition, Study 1 Extension

Notes: Impressions from users with unknown gender (.09%) are not reported in the table for parsimony.

In terms of divergent delivery, the empirical CDFs (see Web Appendix E.2 for details on the approach) lie above the diagonal, indicating an overrepresentation of small p -values and thus the presence of divergent delivery. 58.3% of balance checks (14 out of 24) are significant at the $p < .05$ level. Consistently, Kolmogorov-Smirnov (KS) tests ($p < .001$ for age and $p = .0013$ for gender) and Cramér-von Mises (CvM) tests ($p < .001$ for age and $p = .0067$ for gender) are significant, indicating non-uniformity (Burtch et al., 2025).

In terms of magnitude, 100% of balance checks yield Cramér’s $V < .1$, a threshold conventionally interpreted as indicating small effects (Cohen, 1988). Mean values of Cramér’s V are close to zero for both age ($M = .0105$) and gender ($M = .0108$).

Table A7 presents the results, confirming the significantly higher CTR ($\beta = .1790$, $p < .001$) and ER ($\beta = .1741$, $p < .001$) for AI on TikTok .

¹TikTok Business Help Center (TikTok, 2025); accessed September 26, 2025.

Model:	(1)	(2)
Dep. Var.:	TikTok CTR	TikTok ER
<i>Video ads</i>		
Human (baseline)		
AI	.1790*** (.0532)	.1741*** (.0386)
<i>Controls</i>		
Audience age	Yes	Yes
Audience gender	Yes	Yes
<i>Fixed effects</i>		
Music track	Yes	Yes
<i>Fit statistics</i>		
Observations	90	90
Squared correlation	.8814	.5133
Pseudo R ²	.9815	.9305
BIC	368.6	251.6

Table A7: Regression results, Study 1 Extension

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination.

E Audience Composition and Divergent Delivery

E.1 Audience Composition Across Conditions

This section reports the audience composition for Studies 1-3. Tables A8, A9, and A10 summarize the distribution of impressions across observable demographic groups by condition and platform, separately for each study. Within each study, we report descriptive statistics pooled across campaigns to characterize the audience composition realized across our field deployment. Since divergent delivery bias arises within a test, we compare the pairwise test differences relative to the human ad condition, which we visualize in Figure A3 in Web Appendix E.2, following Burtch et al. (2025).

Condition	TikTok								Meta							
	Impressions	Gender		Age					Impressions	Gender		Age				
		female	male	18-24	25-34	35-44	45-54	55+		female	male	18-24	25-34	35-44	45-54	55+
Overall	4,324,060	54.0%	45.6%	46.5%	16.0%	11.0%	11.3%	14.9%	3,462,006	28.6%	70.1%	13.4%	23.8%	18.3%	17.6%	27.0%
Human	876,939	50.9%	48.7%	50.7%	16.6%	10.1%	9.9%	12.3%	693,497	28.8%	69.9%	13.4%	23.8%	18.2%	17.5%	27.1%
AI	3,447,121	54.8%	44.8%	45.4%	15.8%	11.2%	11.7%	15.5%	2,768,509	28.6%	70.1%	13.4%	23.8%	18.3%	17.6%	26.9%

Table A8: Audience composition, Study 1

Notes: Impressions from users with unknown gender (.8%) or age (.2%) are not reported in the table for parsimony. Data are pooled across campaigns.

Condition	TikTok								Meta							
	Impressions	Gender		Age					Impressions	Gender		Age				
		female	male	18-24	25-34	35-44	45-54	55+		female	male	18-24	25-34	35-44	45-54	55+
Overall	1,329,425	58.1%	41.1%	57.8%	18.5%	7.3%	6.8%	8.7%	862,564	28.8%	69.8%	13.4%	24.3%	20.6%	17.3%	24.3%
Human	314,121	58.0%	41.3%	61.3%	18.7%	6.5%	5.7%	7.0%	141,329	21.7%	76.8%	13.6%	22.4%	19.5%	18.2%	26.3%
AI: expert selected	521,755	58.6%	40.7%	56.5%	18.7%	7.6%	7.2%	9.2%	423,511	29.7%	68.9%	15.0%	26.6%	20.7%	16.0%	21.6%
AI: randomly sampled	493,549	57.8%	41.4%	56.9%	18.3%	7.6%	7.1%	9.1%	297,724	30.9%	67.8%	11.1%	21.8%	21.1%	18.7%	27.3%

Table A9: Audience composition, Study 2

Notes: Impressions from users with unknown gender (1%) or age (.5%) are not reported in the table for parsimony. Data are pooled across campaigns.

Condition	TikTok					Meta				
	Impressions	Gender		Age		Impressions	Gender		Age	
		female	male	18-24	55+		female	male	18-24	55+
Overall	552,018	48.5%	51.5%	58.7%	41.3%	510,214	45.3%	54.7%	51.1%	48.9%
Human	96,616	48.7%	51.3%	70.4%	29.6%	159,339	36.6%	63.4%	58.9%	41.1%
AI: predicted best CTR	226,358	49.4%	50.6%	57.3%	42.7%	139,216	38.1%	61.9%	63.7%	36.3%
AI: predicted worst CTR	229,044	47.5%	52.5%	55.1%	44.9%	211,659	56.5%	43.5%	36.8%	63.2%

Table A10: Audience composition, Study 3

Notes: In Study 3, each campaign is targeted at a unique age-gender segment. Differences in impressions stem from different CPMs (see Table A22), as discussed in Web Appendix K. Data are pooled across campaigns.

E.2 Divergent Delivery

A/B tests on contemporary digital ad platforms are subject to divergent delivery, i.e., systematic differences in how platforms allocate impressions across audiences even when targeting criteria, budgets, and timing are held constant (Boegershausen et al., 2025; Braun and Schwartz, 2024; Braun et al., 2024; Burtch et al., 2025). Because platform algorithms dynamically optimize delivery based on predicted relevance, different assets within the same campaign may reach different audiences (see Figure 2 in Braun et al., 2024 for an intuitive visualization of divergent delivery). If such delivery differences correlate with both treatment assignment and outcomes, they can bias field study findings.

In the following, we provide transparent diagnostic evidence on the presence and magnitude of divergent delivery in our field studies, to the extent it is observable by us. Importantly, we assume that divergent delivery does not systematically bias our findings towards AI-generated assets across campaigns. Because divergent delivery on unobservable user characteristics remains invisible to us, this assumption cannot be tested directly, representing an important limitation of our field design that we acknowledge explicitly and that motivates our “near-exact replication” in the lab (Boegershausen et al., 2025; see Web Appendix M).

Method

Our diagnostics follow Burtch et al. (2025) in assessing divergent delivery in A/B tests. Crucially, divergent delivery arises within a campaign, not across campaigns. Accordingly,

all individual checks are conducted at the campaign level.

First, we define a campaign as a unique platform $\times$ music track combination. Within each campaign, we compare delivery between the human-created baseline asset and each individual AI-generated asset. This yields multiple pairwise comparisons per campaign because each AI asset could be delivered differentially compared to the human baseline.

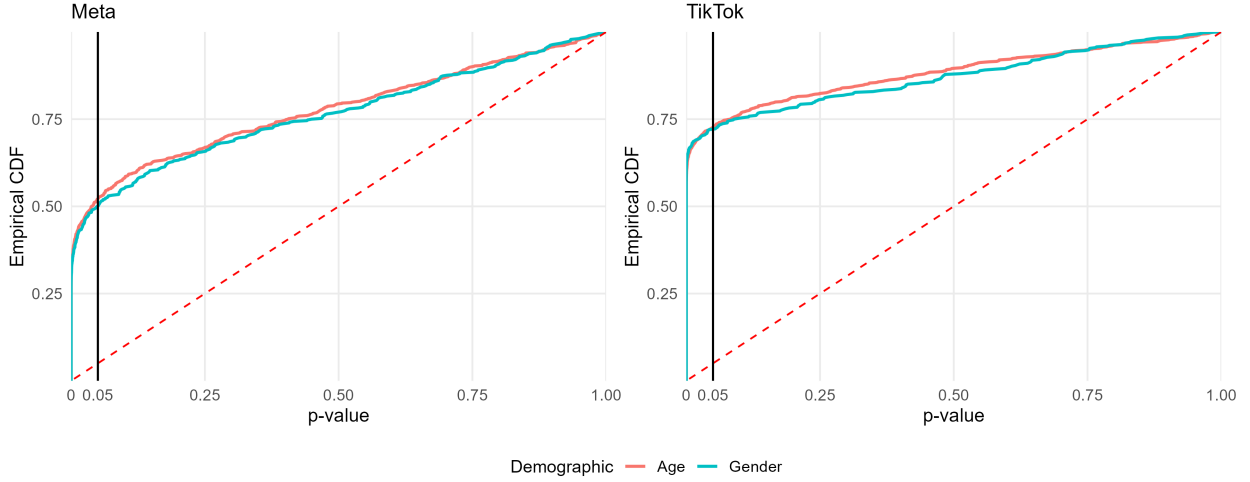
Second, for each asset pair, we test balance on observable demographics using χ^2 tests: for gender (female, male, unknown) and for age (18-24, 25-34, 35-44, 45-54, 55+, unknown). Each test produces a p -value indicating whether delivery differs across observable demographics for that specific within-campaign comparison.

Third, we analyze the distribution of these p -values separately by platform (Meta and TikTok) and observable demographic dimension (age and gender). Specifically, we examine empirical cumulative distribution functions (CDFs) of p -values, as proposed by Burtch et al. (2025). Under perfect randomization and absent systematic delivery differences, p -values should follow a uniform distribution on $[0, 1]$. Deviations from uniformity (especially an excess of small p -values) indicate the presence of divergent delivery. This approach treats balance checks as diagnostic tools: a given comparison either passes or fails a balance test.

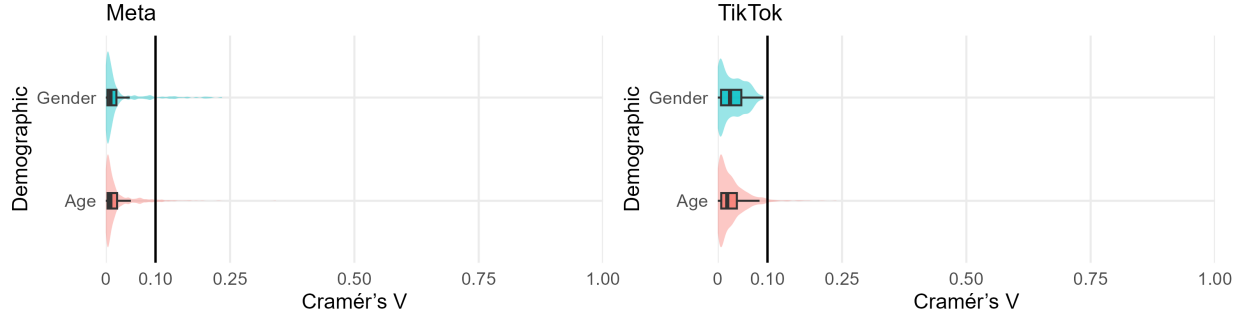
Results

Figure A3a displays the empirical CDFs of p -values from all within-campaign balance checks, separately by platform and demographic dimension. On both platforms, the empirical CDFs lie above the diagonal, indicating an overrepresentation of small p -values and thus the presence of divergent delivery. Approximately 51.4% of balance checks (641 out of 1,248) on Meta and 72.4% on TikTok (1,016 out of 1,404)² are significant at the $p < .05$ level. Consistent with visual inspection of Figure A3a, Kolmogorov-Smirnov tests ($p < .001$) and Cramér-von Mises tests ($p < .001$) are significant, indicating non-uniformity (Burtch et al., 2025). These results demonstrate that divergent delivery is empirically present in our field studies.

²TikTok has more balance checks than Meta because TikTok reports an “unknown” age bucket.



(a) Presence of divergent delivery (empirical CDF of p -values)



(b) Magnitude of divergent delivery (Cramér's V)

Figure A3: Assessment of divergent delivery in terms of presence and magnitude
Notes: The red dashed 45-degree line in panel (a) represents the uniform distribution expected under perfect balance. The vertical black line in panel (a) indicates the significance threshold at $p = .05$, while in panel (b) it marks $V = .1$, below which Cramér's V indicates small effect sizes (Cohen, 1988). Each pairwise test difference is computed within a campaign. We aggregate the resulting p -values across campaigns and studies only to construct the empirical CDF shown in panel (a).

We now turn to effect sizes to separately assess the magnitude of delivery differences, which matters in large-scale campaigns with millions of impressions, where even small deviations yield highly significant test statistics. Figure A3b reports the distribution of Cramér's V for all balance checks (Sella et al., 2021). Across studies, platforms and demographics, 94.6% of Meta's (1,181 out of 1,248) and 98.5% of TikTok's (1,383 out of 1,404) balance checks yield Cramér's $V < .1$, conventionally interpreted as a small effect (Cohen, 1988), with mean values close to zero for both age ($M = .0218$ on Meta, $M = .0276$ on TikTok) and gender ($M = .0263$ on Meta, $M = .0291$ on TikTok).

F Model-Free Evidence

This section reports descriptive comparisons of AI and human video ads as model-free evidence. Table A11 summarizes raw click-through rates, engagement rates, percentage uplifts, and associated χ^2 tests across different aggregation levels by study and platform.

Aggregation:	Model: Platform Dep. Var.:	(1) Overall CTR	(2) TikTok CTR	(3) Meta CTR	(4) Overall ER	(5) TikTok ER	(6) Meta ER
Pooled Studies 1-3	Human	.0918%	.1207%	.0544%	4.8977%	3.7677%	6.3612%
	AI	.1065%	.1509%	.0497%	5.4289%	4.5699%	6.5288%
	Uplift AI vs. Human	+16.0%	+25.0%	-8.7%	+10.8%	+21.3%	+2.6%
	χ^2	37.7***	64.2***	3.3†	1014***	1560***	36.5***
Study 1 only	Human	.0928%	.1208%	.0574%	5.1828%	3.9806%	6.7031%
	AI	.1082%	.1529%	.0526%	5.1107%	4.2828%	6.1415%
	Uplift AI vs. Human	+16.7%	+26.6%	-8.4%	-1.4%	+7.6%	-8.4%
	χ^2	28.4***	49.2***	2.3	13.4***	158***	298***
Study 2 only	Human	.0863%	.1060%	.0425%	4.2753%	3.0001%	7.1097%
	AI: randomly sampled (RS)	.1001%	.1311%	.0487%	5.7553%	3.9120%	8.7977%
	AI: expert selected (ES)	.0843%	.1217%	.0383%	5.3798%	3.8206%	7.3006%
	Uplift RS vs. Human	+16.0%	+23.7%	+14.7%	+34.6%	+30.7%	23.7%
	χ^2	5.7*	9.8**	.7	1281***	473***	360***
	Uplift ES vs. Human	-2.3%	+14.8%	-9.9%	+25.8%	+27.3%	+2.7%
	χ^2	.1	4.0*	.4	786***	389***	5.7*
	Uplift RS vs. ES	+18.7%	+7.7%	+27.3%	+7.0%	+2.6%	+20.5%
Study 3 only	Human	.0957%	.1677%	.0521%	4.2554%	4.3316%	4.2093%
	AI: predicted best CTR (PB)	.1480%	.2090%	.0488%	8.7099%	10.8540%	5.2235%
	AI: predicted worst CTR (PW)	.1071%	.1720%	.0369%	6.7152%	5.7884%	7.7181%
	Uplift PB vs. Human	+54.6%	+24.6%	-6.2%	+104.7%	+150.6%	+24.1%
	χ^2	32.1***	5.7*	.1	4665***	3551***	171***
	Uplift PW vs. Human	+11.9%	+2.6%	-29.3%	+57.8%	+33.6%	+83.4%
	χ^2	1.9	.1	4.5*	1789***	284***	1921***
	Uplift PB vs. PW	+38.2%	+21.5%	+32.5%	+29.7%	+87.5%	-32.3%
	χ^2	26.3***	8.0**	2.6	1129***	3835***	832***

Table A11: Model-free evidence

Notes: Each reported CTR and ER value is calculated by aggregating impressions, clicks, and engagement within the corresponding condition, platform, and study grouping shown in the rows and columns. The *pooled* rows aggregate observations across Studies 1-3, whereas the study-specific rows report statistics separately for each study. Columns (1) and (4) aggregate *overall* across platforms, while the remaining columns report platform-specific statistics. χ^2 tests compare the corresponding AI and human conditions within each aggregation level. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

G Meta Platform Split: Instagram vs. Facebook

This section separates the results for Meta into Instagram and Facebook and reports them in Tables A12, A13, and A14.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Instagram	Facebook	Meta	Instagram	Facebook	Meta
	CTR	CTR	CTR	ER	ER	ER
<i>Video ads</i>						
Human (baseline)						
AI	.0358 (.0767)	-.2190* (.1096)	-.2178* (.1047)	-.0456 (.0446)	-.1042* (.0438)	-.1042** (.0442)
Facebook (baseline)						
AI $\times$ Instagram			.2541 $\dagger$ (.1388)			.0587 (.0610)
<i>Controls</i>						
Audience age	No	No	No	No	No	No
Audience gender	No	No	No	No	No	No
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Meta platform	No	No	Yes	No	No	Yes
<i>Fit statistics</i>						
Observations	115	115	230	115	115	230
Squared correlation	.3906	.5780	.4294	.8488	.6816	.8873
Pseudo R ²	.9669	.9606	.9613	.9965	.9957	.9954
BIC	637	650	1,283	1,587	4,691	7,069

Table A12: Regression results for Meta platform split into Instagram vs. Facebook, Study 1

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, $\dagger$ $p < .1$.

The unit of observation is an ad-platform combination as the age and gender splits are not available, when splitting the Meta results by Instagram and Facebook.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Instagram CTR	Facebook CTR	Meta CTR	Instagram ER	Facebook ER	Meta ER
<i>Video ads</i>						
Human (baseline)						
AI: randomly sampled (RS)	.1347 (.1500)	.4678* (.2022)	.4367* (.1889)	.0531 (.0728)	.0278 (.0387)	.0233 (.0345)
AI: expert selected (ES)	-.2147† (.1239)	.1411 (.2063)	.0587 (.2135)	-.0855 (.0783)	.0272 (.0307)	.0233 (.0345)
Facebook (baseline)						
RS × Instagram			-.3308 (.2526)			.0107 (.0698)
ES × Instagram			-.2181 (.2687)			-.0553 (.0852)
<i>Controls</i>						
Audience age	No	No	No	No	No	No
Audience gender	No	No	No	No	No	No
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Meta platform	No	No	Yes	No	No	Yes
<i>Fit statistics</i>						
Observations	63	63	126	63	63	126
Squared correlation	.0518	.3333	.2604	.3800	.7639	.8630
Pseudo R ²	.9433	.9397	.9402	.9925	.9966	.9954
BIC	233	212	443	770	1,278	2,237

Table A13: Regression results for Meta platform split into Instagram vs. Facebook, Study 2

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

The unit of observation is an ad-platform combination as the age and gender splits are not available, when splitting the Meta results by Instagram and Facebook.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Instagram CTR	Facebook CTR	Meta CTR	Instagram ER	Facebook ER	Meta ER
<i>Video ads</i>						
Human (baseline)						
AI: predicted best CTR (PB)	.1302 (.1584)	-.5396 (.4416)	-.5624 (.4312)	.2442*** (.0643)	.2857 (.1924)	.3014† (.1804)
AI: predicted worst CTR (PW)	-.0553 (.1880)	-.3935** (.1488)	-.4170** (.1517)	.2672*** (.0706)	.3668** (.1281)	.3898** (.1242)
Facebook (baseline)						
PB × Instagram			.6769 (.4594)			-.0513 (.1999)
PW × Instagram			.3352 (.2366)			-.1456 (.1496)
<i>Controls</i>						
Audience age	No	No	No	No	No	No
Audience gender	No	No	No	No	No	No
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Meta platform	No	No	Yes	No	No	Yes
<i>Fit statistics</i>						
Observations	36	36	72	36	36	72
Squared correlation	.0006	.3386	.2215	.8338	.6927	.7136
Pseudo R ²	.9574	.9561	.9554	.9941	.9900	.9911
BIC	132	76	216	418	1,550	2,096

Table A14: Regression results for Meta platform split into Instagram vs. Facebook, Study 3

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

The unit of observation is an ad-platform combination as the age and gender splits are not available, when splitting the Meta results by Instagram and Facebook.

H Waves

Study 1 was run in multiple waves. Table A15 reports the regression results for Study 1 with wave fixed effects instead of music track fixed effects.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall	TikTok	Meta	Overall	TikTok	Meta
	CTR	CTR	CTR	ER	ER	ER
<i>Video ads</i>						
Human (baseline)						
AI	.1730*** (.0364)	.2574*** (.0411)	-.0869 (.0704)	-.0435* (.0200)	.0199 (.0259)	-.0948*** (.0281)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Wave	Yes	Yes	Yes	Yes	Yes	Yes
Music track	No	No	No	No	No	No
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	4,114	2,239	1,875	4,114	2,239	1,875
Squared correlation	.0223	.0126	.0070	.1499	.0804	.4776
Pseudo R ²	.9248	.9396	.8786	.9809	.9825	.9804
BIC	9,766	5,841	3,915	60,157	26,535	31,824

Table A15: Regression results including wave controls, Study 1

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination.

I Expert Selection

In Study 2, one industry expert did not perform significantly better at selecting four out of eight video ads for CTR than random sampling. To better understand whether this difficulty in predicting ad performance generalizes across populations and task formats, we conducted five additional selection studies (see Table 3 in Section 5.2). These studies used the 28 AI-generated video ads from the seven campaigns in Study 2 (four ads per campaign from the ‘AI: randomly sampled’ condition). Critically, we know the actual field performance for each ad, allowing us to assess selection accuracy against ground-truth CTR. Table A16 shows the CTR differences across the four ads within each campaign. The winning ad outperformed the second best ad by 27.5% on average (ranging from 3.6% to 85.9% across campaigns), indicating meaningful performance variation that, in principle, could be detected.

Campaign	Performance of AI Video Ads			
	Best	vs. 2nd	vs. 3rd	vs. 4th
	CTR	Uplift (CTR)	Uplift (CTR)	Uplift (CTR)
1	.1807%	+29.7% (.1393%)	+43.3% (.1261%)	+61.7% (.1117%)
2	.1047%	+15.9% (.0904%)	+46.8% (.0713%)	+73.5% (.0604%)
3	.1146%	+5.8% (.1083%)	+10.1% (.1041%)	+66.7% (.0688%)
4	.1315%	+85.9% (.0707%)	+161.3% (.0503%)	+170.8% (.0485%)
5	.1047%	+3.6% (.1011%)	+17.4% (.0892%)	+36.2% (.0769%)
6	.1810%	+39.5% (.1297%)	+48.9% (.1216%)	+101.5% (.0898%)
7	.1306%	+43.9% (.0908%)	+53.0% (.0854%)	+63.9% (.0797%)
Overall	.1327%	+27.5% (.1040%)	+40.6% (.0944%)	+82.2% (.0728%)

Table A16: CTR differences for assets used in the selection studies

I.1 Selection Studies I and II: Marketing Professionals and Laypeople

To test whether the difficulty in predicting ad performance extends beyond the industry expert from Study 2, we ran two preregistered lab studies with marketing professionals and laypeople, using the same ads and selection setup as described above (AsPredicted #238172 and AsPredicted #237772). Specifically, we recruited $n = 98$ marketing professionals (two were excluded due to failed attention checks) and $n = 99$ laypeople (one was excluded due to a failed attention check). Marketing professionals were pre-screened on CloudResearch

for the occupation “Marketing and Sales,” while laypeople were not. After passing the same attention checks as in previous lab studies, respondents watched the four 10-second video ads for every music track and answered the questions detailed in Table A17. To minimize response bias related to preconceptions about AI-generated content, we deliberately avoided mentioning any connection to AI in the entire survey.

Items	Response Options
<i>Based on your experience, which video ad will perform best as a social media ad (e.g., on TikTok, Instagram or Facebook) in terms of click-through rate (CTR)?</i>	Video 1 (top-left) Video 2 (top-right) Video 3 (bottom-left) Video 4 (bottom-right)
<i>Please indicate your gender</i>	Male Female Non-binary Prefer not to say
<i>Please indicate your age</i>	Numerical input
<i>How frequently do you use social media (e.g., TikTok, Instagram or Facebook)?</i>	Not at all Once per month Once per week Once per day Multiple times per day
<i>How many years of professional experience do you have?</i>	Numerical input
<i>Do you work in a marketing-related job?</i>	Yes No Prefer not to say

Table A17: Overview of scales used for Selection Studies I and II

Results show that neither marketing professionals nor laypeople outperform the chance baseline of 25% in selecting the best performing video ad, reaching accuracies of 23.6% ($t(98) = -.4674$, $p = .6413$) and 24.2% ($t(97) = -.9738$, $p = .3326$), respectively. Notably, there is no significant difference in accuracy between laypeople and professionals ($\beta = -.0353$, $p = .2075$), see model (1) in Table A18.

Does experience improve prediction accuracy? As expected, we find that professionals’ accuracy increases with experience, although this effect is only marginally significant ($\beta_{Professionals \times Experience_z} = .0435$, $p = .0672$). Consistent with this finding, the interaction with age is significant ($\tilde{\beta}_{Professionals \times Age_z} = .0606$, $p = .0073$). However, even the ten most experienced professionals in our sample achieve only an average accuracy of 27.1%

($t(9) = .5417$, $p = .6012$). In other words, more experienced, older professionals show slightly better forecasting ability, though this advantage remains modest and statistically indistinguishable from chance. Importantly, this pattern diverges from the field finding that older users show lower responsiveness to AI assets, which reflects their own behavior rather than their ability to forecast what others will click.

Model:	(1)	(2)
Dep. Var.:	Overall Accuracy	Overall Accuracy
Constant	.2393*** (.0300)	.2371*** (.0290)
Experience _z	-.0212 (.0186)	-.0362† (.0197)
Professionals	-.0353 (.0279)	-.0359 (.0272)
Professionals × Experience _z		.0435† (.0237)
<i>Controls</i>		
Age	Yes	Yes
Gender	Yes	Yes
Social Media Usage	Yes	Yes
Marketing Job	Yes	Yes
<i>Fit statistics</i>		
Observations	197	197
R ²	.0355	.0539
Adjusted R ²	-.0109	.0030

Table A18: Regression results, laypeople and professionals asset selection

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. Experience_z is standardized ($M_{Experience} = 15.05$, $SD = 10.50$). For laypeople, experience refers to self-reported years of overall professional experience, regardless of occupation.

I.2 Selection Study III: Industry Experts

We explicitly instructed $n = 4$ industry experts to select the best-performing asset in terms of CTR.³ After watching the four AI video ads per music track, they selected one winning asset. Notably, none of the industry experts predicted the best performing ad for any of the music tracks, resulting in an accuracy of 0% (0 out of 28 predictions) compared to a 25% chance baseline. Furthermore, we find a low inter-rater reliability among the four industry experts ($\kappa_{Fleiss} = .111$, $p = .15$), indicating limited agreement on selection criteria despite being explicitly prompted to optimize for CTR.

Given the small sample size, we interpret this 0% accuracy as consistent with the null finding from Selection Studies I and II rather than evidence that industry experts perform

³As campaigns for catalog tracks typically use platform-agnostic creatives, we instructed the experts to select the ad with the highest average CTR across both platforms, reflecting their real-world practice.

systematically worse than chance. Importantly, whether selection studies show at-chance or below-chance performance, the substantive conclusion remains: expert selection does not improve outcomes over random sampling from MAGE, as confirmed by the field evidence in Study 2. The low inter-rater agreement suggests experts may apply different implicit criteria such as brand safety, visual coherence, or artist fit.

I.3 Selection Study IV: Rank Ordering

To provide a more sensitive test of predictive ability, we conducted an additional preregistered selection study ($n = 99$ complete responses, AsPredicted #268022), asking a separate group of marketing professionals to rank the four video ads per campaign for one specified target segment (20 target segments with ~ 5 professionals each), instead of predicting the single best video ad.⁴ After watching the video ads, we asked *“Based on your experience, how would you rank all four video ads’ click-through rates (CTR) on {platform} among {gender} users aged {age}?”*. This setup allows us to compare both (i) success rates in picking the best video ad and (ii) rank correlation at the campaign $\times$ target segment level.

As the AI model can actually select the best video ad per target segment (e.g., TikTok, male, 18-24), we compare both professionals and the AI model to realized performance. Both professionals (Spearman’s $\rho = .0871$, $p = .0861$) and the AI model (Spearman’s $\rho = .1786$, $p < .001$) exhibit positive rank correlation with realized performance, but the AI model performs significantly better and the human correlation is only marginally significant. Consistently, the success rates show that the AI model performs better than humans (30% vs. 24.8%), while humans again do not outperform the random chance baseline of 25% ($t(98) = -0.0992$, $p = .5394$). Even with 99 professionals generating 2,772 total judgments across campaigns, the ordinal association with field CTR remains modest, suggesting that sample size alone does not fully resolve the prediction challenge for humans. Regressing the ads’ ranks on their visual features reveals correlated coefficient patterns between professionals and the evaluative AI model ($r = .7783$, $p < .001$; Table A19). However, the AI’s superior

⁴We excluded all professionals who participated in Selection Study I.

predictive performance suggests it better captures complex feature interactions that human judgment cannot reliably predict.

Model:	(1)	(2)
Dep. Var.:	Human Inverse Rank	Evaluative AI Inverse Rank
<i>Narrative:</i>		
Intro	-1.259** (.4393)	-4.151*** (.7993)
Outro	.5342** (.1796)	1.879*** (.3316)
<i>Content:</i>		
Shiny	-.9513*** (.2451)	-1.038* (.4812)
Product	-.3089 (.2878)	.6035 (.5376)
Reflection	.4437* (.1933)	.1172 (.3371)
Fly-in	.4764* (.2196)	1.473*** (.3073)
Zoom	-.0457 (.1390)	1.371*** (.2999)
Waveform	1.134*** (.2921)	2.703*** (.5311)
<i>Background:</i>		
GenAI Image	.0576 (.3148)	-1.375** (.5108)
GenAI Images	.7211** (.2427)	.1371 (.5502)
GenAI Split Images	.5436 (.3478)	.6823 (.6049)
GenAI Videos	.0430 (.1951)	-.3803 (.3455)
GenAI 2× Videos	-.2355† (.1231)	-.6745*** (.2010)
<i>Text:</i>		
Position: Below	.1043 (.1166)	.8339*** (.2155)
Font: Medium	-.1408 (.0947)	.4415* (.1724)
Font: Small	.2282 (.1415)	.2570 (.2275)
<i>Audio:</i>		
Louder	-.1096 (.0861)	.5709*** (.1719)
Speed: Fast @105%	-.0304 (.1108)	-.4179 (.1976)
<i>Fixed effects</i>		
Respondent	Yes	No
Segment age	Yes	Yes
Segment gender	Yes	Yes
Segment platform	Yes	Yes
Music track	Yes	Yes
<i>Fit statistics</i>		
Observations: ^a	2,772	560
R ²	.0260	.3308

Table A19: Regression results for segment-level CTR ranking of (1) human professionals and (2) the evaluative AI model

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

^a Observation counts between models (1) and (2) differ as, on average, five respondents rank the video ads for every segment, while the evaluative AI ranks the assets deterministically once per segment. An ordinal regression yields substantively consistent results to this OLS regression.

I.4 Selection Study V: Selection by Engagement

But can humans not follow the briefing for a specified target because they have some other unobservable selection criteria that overrule the prompted target? Then we would expect the expert to make consistent selections regardless of the instructed metric. To explore this possibility, we asked the same industry expert who had previously selected the four out of eight assets for Study 2, instructed to select based on CTR (see Section 5), to repeat the task more than half a year later with the same assets, but this time selecting assets explicitly for a different metric, namely, engagement.

Out of the 28 selections made, only 16 assets (57.1%) overlap with the CTR-based choices, marginally above random chance (expected: 14/28) and not statistically significant ($\chi^2(1, N = 28) = .5714, p = .4497$). This indicates that the expert did adjust selections based on the instructed metric. Importantly, random selection can be ruled out as an alternative explanation, as subsequent analysis reveals that the expert exhibits systematic preferences for specific design parameters regardless of the instructed metric.

Follow-up conversations with experts from the industry partner suggest that they select differentially based on campaign targets, yet cannot fully suppress aesthetic biases related to visual appeal and brand coherence, which are critical considerations given the importance of artist approval for assets that must remain ‘on brand’ (see Hou et al. (2025) for related work on cognitive fixation among designers). As one expert noted: “We are always looking for elements that are on brand with the artist, visually appealing, and aesthetically in line with the overall campaign.”

To analyze how the AI assets’ design parameters relate to human selection decisions, we use logistic regression with expert selection (asset selected vs. not selected) as the dependent variable, as detailed in Table A20 (animation type, which is one of the most salient design parameters of MAGE) and Table A21 (all remaining parameters). Both tables contrast expert selection preferences in models (1) and (5) for CTR and ER, respectively, with real-world

performance pooled across Studies 1 and 2⁵ in models (2) through (4) and (6) through (8), surfacing aesthetic biases and guardrails. For instance, the expert systematically deselects the ‘shiny’ animation, both when instructed to select based on CTR ($\beta = -2.414, p = .0089$) or engagement ($\beta = -29.15, p < .001$), even though it is significantly positively associated with real-world CTR ($\beta = .1556, p = .0107$) and insignificantly related to engagement ($\beta = .0303, p = .2464$ on TikTok and $\beta = -.0041, p = .9219$ on Meta). Sharing such insights could help “teach” experts the target function of the evaluative AI model and improve their ability to evaluate AI-generated ads.

Model:	(1) Expert Selection	(2) Overall (Pooled)	(3) TikTok (Pooled)	(4) Meta (Pooled)	(5) Expert Selection	(6) Overall (Pooled)	(7) TikTok (Pooled)	(8) Meta (Pooled)
Dep. Var.:	CTR	CTR	CTR	CTR	ER	ER	ER	ER
No animation (baseline)								
Shiny	-2.414** (.9225)	.1556* (.0609)	.1682** (.0612)	.1216 (.1355)	-29.15*** (5.920)	-.0004 (.0306)	.0303 (.0261)	-.0041 (.0420)
Product	-2.575*** (.9220)	-.0424 (.1359)	.1199 (.1408)	-.1160 (.3562)	-23.77† (14.07)	-.1794* (.0731)	-.2332** (.0727)	-.2419* (.1070)
Reflection	-1.620† (.9614)	.0302 (.0527)	.0188 (.0565)	.0438 (.1054)	31.95 (25.65)	.0692* (.0290)	.0690** (.0267)	.0490 (.0367)
Fly-in	-.0916 (.9251)	-.0762† (.0640)	-.0635 (.0679)	-.0147 (.1291)	24.75 (17.21)	.0233 (.0311)	.0466 (.0327)	-.0023 (.0421)
Zoom	-1.149 (.7448)	.0255 (.0517)	.0244 (.0561)	-.0282 (.1189)	12.79 (9.299)	-.0046 (.0285)	.0217 (.0289)	-.0149 (.0393)
Boom	-2.411** (.9145)	-.0352 (.0934)	-.0488 (.0988)	.0248 (.1810)	- (.0460)	.0666 (.0460)	.0427 (.0486)	.0906 (.0625)
Shear	-2.963† (1.636)	-.0898 (.1598)	-.2051 (.1786)	.2487 (.3344)	- (.0955)	.1539 (.0955)	.3060** (.0959)	.0089 (.1121)
<i>Controls</i>								
Visual ^a	Yes	Yes	Yes	Yes	Yes ^b	Yes	Yes	Yes
Audio ^a	Yes	Yes	Yes	Yes	Yes ^b	Yes	Yes	Yes
Audience age	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Audience gender	No	Yes	Yes	Yes	No	Yes	Yes	Yes
<i>Fixed effects</i>								
Music track	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Platform	No	Yes	No	No	No	Yes	No	No
<i>Fit statistics</i>								
Observations:	256	5,063	2,724	2,339	56 ^b	5,063	2,724	2,339
Squared correlation	.2967	.0130	.0073	.0122	.7170	.1562	.0927	.3472
Pseudo R ²	.2378	.9228	.9386	.8772	.6687	.9835	.9867	.9841
BIC	653	10,753	6,658	4,300	126	54,243	21,275	21,289

Table A20: Regression results comparing expert selection with real-world performance

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

^a A list of all visual and audio controls can be found in Table A21 in Web Appendix I.

^b Model (5) is based on expert selection data from Study 2 in which not all feature variants are present in the available AI video ads. All other models with expert selection and field data from Studies 1 and 2.

⁵Study 3 is based on AI selection (instead of expert selection) and therefore not included in the table.

Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Expert	Overall	TikTok	Meta	Expert	Overall	TikTok	Meta
Dep. Var.:	Selection	(Pooled)	(Pooled)	(Pooled)	Selection	(Pooled)	(Pooled)	(Pooled)
	CTR	CTR	CTR	CTR	ER	ER	ER	ER
<i>Narrative:</i>								
Intro	1.758* (.7809)	.4381*** (.0573)	.6006*** (.0522)	-.4105** (.1430)	92.77*** (27.19)	.0006 (.0355)	-.0388 (.0408)	.0543 (.0494)
Outro	.7888 (1.173)	.0644 (.0984)	.0573 (.1034)	-.0461 (.1827)	39.85* (17.22)	.0429 (.0436)	.1026* (.0522)	-.0272 (.0589)
<i>Content:</i>								
Position: Middle	.1401 (.7877)	.0184 (.0526)	.0014 (.0597)	.0718 (.1072)	- -	-.0158 (.0263)	.0314 (.0252)	-.0405 (.0351)
Position: Top	-.0775 (.7100)	.0481 (.0560)	.0291 (.0604)	.1178 (.1297)	- -	-.0513† (.0295)	-.0360 (.0277)	-.0458 (.0404)
Size: Medium	-.1021 (.4986)	-.0134 (.0382)	-.0313 (.0406)	.0075 (.0853)	- -	.0029 (.0213)	.0206 (.0215)	-.0014 (.0284)
Size: Small	-.4616 (.7165)	-.0265 (.0542)	-.0554 (.0604)	.0655 (.1060)	- -	-.0678* (.0308)	-.1081*** (.0293)	-.0393 (.0417)
Edge radius: Small	.4443 (.5508)	-.0653 (.0427)	-.0751† (.0446)	-.0016 (.0901)	- -	-.0190 (.0231)	-.0092 (.0205)	-.0553† (.0309)
Edge radius: Large	.8793† (.4958)	-.0455 (.0446)	-.0619 (.0466)	.0019 (.0956)	- -	-.0053 (.0256)	.0151 (.0239)	-.0321 (.0373)
Shadow	.4906 (.4213)	.0798† (.0430)	.0707 (.0442)	.0887 (.0919)	- -	.0225 (.0230)	-.0059 (.0225)	-.0411 (.0300)
Waveform	-.0203 (.4424)	-.0675† (.0358)	-.0664† (.0373)	-.0547 (.0751)	32.34 (20.49)	.0173 (.0192)	-.0067 (.0182)	.0428† (.0254)
<i>Background:</i>								
Color	-1.188 (2.111)	.1757 (.1967)	.4292* (.2146)	-.6555 (.4005)	- -	-.0178 (.1042)	.0089 (.0941)	-.0373 (.1272)
Packshot	.3303 (1.027)	.0597 (.0956)	.0953 (.0977)	-.1887 (.2027)	19.71 (16.91)	.0687 (.0523)	.0547 (.0565)	.0695 (.0742)
GenAI Image	1.618 (1.028)	.0793 (.1046)	.1750† (.1041)	-.3589 (.2399)	-13.89* (6.193)	-.0658 (.0546)	-.0105 (.0575)	-.1175 (.0860)
GenAI Images	.8135 (1.893)	.2237 (.1792)	.4743* (.1847)	-.8838* (.3813)	14.41 (15.07)	-.0308 (.0831)	.0471 (.0849)	-.0534 (.1265)
GenAI Videos	.1594 (1.779)	.2842† (.1643)	.5285** (.1713)	-.7146* (.3439)	2.500 (7.361)	.0737 (.0787)	.1237 (.0849)	.0537 (.1223)
GenAI 2× Videos	1.454 (1.970)	.4287* (.1812)	.7051*** (.1901)	-.7793† (.4052)	-12.51* (5.648)	-.0047 (.0899)	.0280 (.0887)	-.0250 (.1393)
Animation: Move	.2773 (.7111)	.2042** (.0793)	.3139*** (.0851)	-.1026 (.1694)	- -	.0407 (.0355)	.0866* (.0348)	-.0041 (.0583)
Animation: Multiple	1.347 (1.314)	-.0861 (.0899)	-.1683† (.1012)	.2902 (.1856)	- -	.0617 (.0464)	.0535 (.0410)	.0327 (.0612)
Motion Bucket ID	.0017 (.0032)	-.0002 (.0002)	-.0002 (.0003)	.0003 (.0005)	- -	-.0003* (.0001)	-.0002† (.0001)	-.0003* (.0002)
Brightness: Light	-.8159 (.5321)	-.1381*** (.0400)	-.1906*** (.0424)	.0654 (.0846)	- -	-.0835*** (.0218)	-.0937*** (.0215)	-.0785** (.0298)
<i>Text:</i>								
Position: Below	-.2308 (.4572)	.0564 (.0399)	.0790† (.0428)	-.0295 (.0826)	4.156* (1.984)	-.0254 (.0236)	-.0354 (.0224)	-.0112 (.0314)
Font: Medium	-.3194 (.5581)	.0218 (.0378)	.0426 (.0408)	-.1078 (.0799)	16.69 (10.47)	-.0149 (.0215)	-.0088 (.0203)	-.0122 (.0284)
Font: Small	.4098 (.5143)	.0068 (.0401)	.0052 (.0432)	.0115 (.0766)	13.15 (9.446)	-.0086 (.0208)	-.0102 (.0201)	.0008 (.0271)
Color: Custom	.7881 (.5224)	-.0131 (.0399)	-.0226 (.0432)	.0413 (.0844)	- -	-.0542** (.0194)	-.0970*** (.0202)	-.0217 (.0262)
Color: White	-.2196 (.5578)	.0363 (.0533)	.0359 (.0561)	.2108† (.1157)	- -	-.0273 (.0283)	-.0637* (.0285)	-.0092 (.0375)
Animation	-3.138* (1.444)	-.0438 (.1190)	-.0309 (.1347)	-.0571 (.3049)	9.453 (6.381)	.0811 (.0599)	.1041† (.0614)	.0273 (.0933)
<i>Audio:</i>								
Louder	-.2735 (.4109)	-.0048 (.0332)	.0080 (.0358)	-.0943 (.0685)	-15.72 (11.38)	-.0237 (.0173)	-.0040 (.0174)	-.0358 (.0230)
Speed: Slow @90%	-.8890 (.6076)	.0308 (.0495)	.0881† (.0517)	-.1814† (.1013)	- -	-.0391 (.0268)	-.0201 (.0271)	-.0550 (.0348)
Speed: Fast @105%	-.6022 (.8907)	-.0831 (.0685)	-.0592 (.0748)	-.1902 (.1559)	- -	-.1037* (.0410)	-.0799* (.0326)	-.0582 (.0619)
Speed: Fast @110%	-1.371* (.6136)	.0343 (.0490)	.0731 (.0538)	-.0679 (.1043)	- -	-.0019 (.0251)	-.0062 (.0239)	.0027 (.0327)
<i>Controls</i>								
Animation ^a	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Audience age	No	Yes	Yes	Yes	No	Yes	Yes	Yes
Audience gender	No	Yes	Yes	Yes	No	Yes	Yes	Yes
<i>Fixed effects</i>								
Music track	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Platform	No	Yes	No	No	No	Yes	No	No
<i>Fit statistics</i>								
Observations:	256	5,063	2,724	2,339	56	5,063	2,724	2,339
Squared Cor.	.2967	.0130	.0073	.0122	.7170	.1562	.0927	.3472
Pseudo R ²	.2378	.9228	.9386	.8772	.6687	.9835	.9867	.9841
BIC	653	10,753	6,658	4,300	126	54,243	21,275	21,289

Table A21: Regression results, expert selection and field performance

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

^a The animation results can be found in Table A20. Model (5) is based on expert selection data from Study 2. All other models with expert selection and field data from Studies 1 and 2.

J Segment-Specific Content Creation

J.1 Training an Evaluative AI Model for AI Video Selection

To train Φ , we use a grid search and determine hyperparameters, which results in a learning rate of .01, a maximum depth of 3, a subsample of .8, and 200 estimators.

J.2 Applying the Evaluative AI Model for AI Video Selection

The selection of the AI assets for the eight campaign variants relies on platform (TikTok and Meta), gender (female and male), and two age bins, namely, 18-24 (younger audience) and 55+ (older audience), see Figure A4. We generate a total of eight variants per segment, that is, four predicted best and four predicted worst.

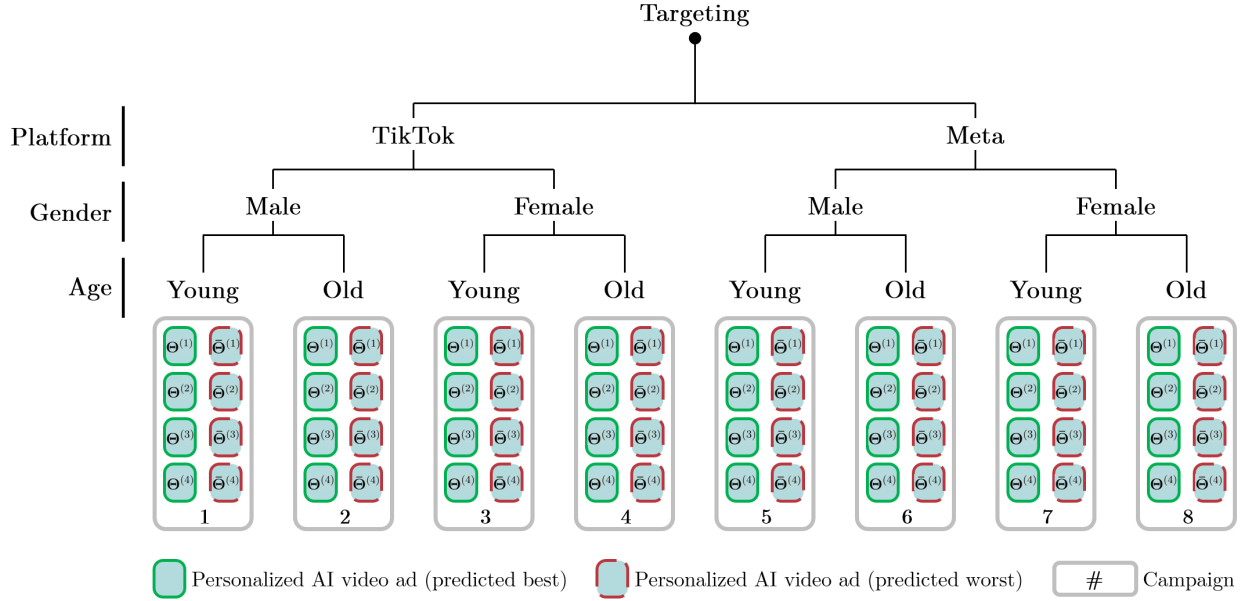


Figure A4: Segment-specific content for target groups, Study 3

Notes: Every campaign also includes one human video ad; it is the same one for all eight campaign variants.

K Cost Per Mille (CPM)

This section reports CPM differences across target segments in Study 3. Table A22 summarizes average CPM by platform, age, gender, and condition, while Table A23 reports regression results for CPM differences across conditions. Because the cost per mille (CPM) is lower for younger than older audiences, we observe more impressions for the former. Importantly, the CPM does not differ significantly between the human video ads and the AI video ads overall ($\tilde{\beta} = -.0262$, $p = .7234$).

Age:	18-24		55+	
Gender:	Female	Male	Female	Male
	CPM	CPM	CPM	CPM
TikTok	1.5	1.6	2.4	2.0
Human	1.5	1.5	2.4	2.0
AI: predicted best CTR	1.5	1.6	2.4	2.0
AI: predicted worst CTR	1.6	1.6	2.4	2.0
Meta	2.0	1.8	2.3	1.8
Human	2.1	1.8	2.2	1.7
AI: predicted best CTR	2.0	1.8	2.3	1.8
AI: predicted worst CTR	2.0	1.8	2.3	1.8

Table A22: Average CPM for target segments, Study 3

Notes: CPM in GBP.

Model:	(1)	(2)	(3)
Dep. Var.:	Overall CPM	TikTok CPM	Meta CPM
<i>Video ads</i>			
Human (baseline)			
AI: predicted best CTR	.0588 (.0910)	.0423 (.0526)	.0429 (.0446)
AI: predicted worst CTR	.0492 (.0926)	.0590 (.0515)	.0456 (.0442)
<i>Controls</i>			
Audience age: 55+ (vs. 18-24)	.3750*** (.0548)	.6486*** (.0344)	.0856* (.0359)
Audience gender: Male (vs. Female)	-.2434*** (.0526)	-.1338*** (.0324)	-.3651*** (.0326)
<i>Fixed effects</i>			
Platform	Yes	No	No
<i>Fit statistics</i>			
Observations	72	36	36
R ²	.6458	.9281	.8887
Within R ²	.6225	—	—

Table A23: OLS regression results for CPM, Study 3

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination.

L Pooled Results Across Studies 1-3

L.1 Ad-level Results

Figure A5 complements the pooled regressions in Section 7 by presenting the underlying ad-level AI estimates across Studies 1-3. Following Gordon et al. (2023), Figure A5’s four panels display the full distribution of individual treatment effects, with one panel for each combination of outcome (CTR and ER) and platform (Meta and TikTok). Each point reports the estimated coefficient for a single AI-generated ad relative to its matched human baseline within the same campaign, and each whisker shows the associated 95% confidence interval. Within every panel, the assets are ordered by their estimated uplift, and color denotes the underlying study and condition.

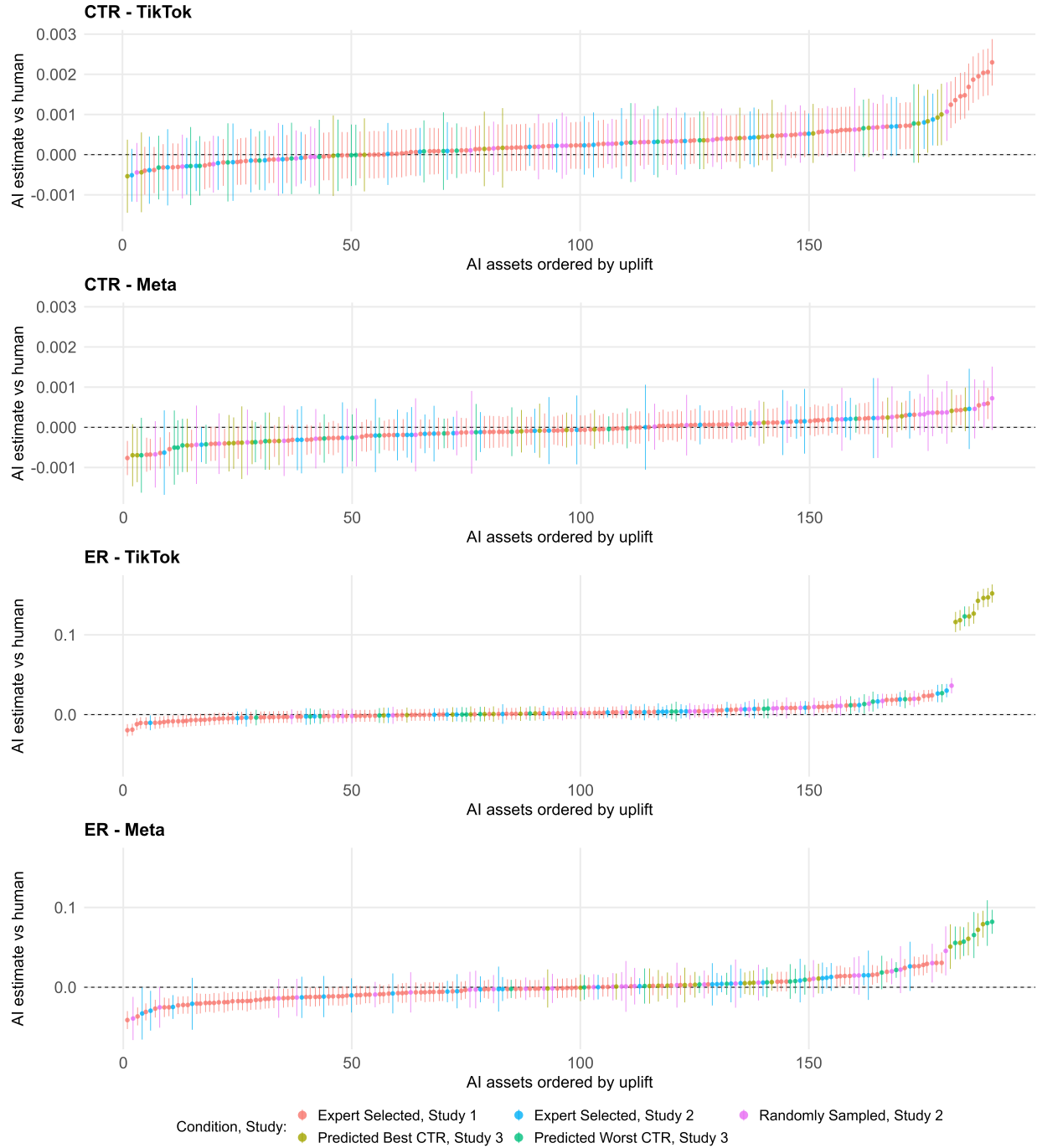


Figure A5: Ad-level AI estimates

Notes: The AI estimates are obtained from the regressions estimated on the data from all studies. Each estimate reflects a single AI asset. Positive estimates indicate that the AI-generated ad outperformed the matched human benchmark on the respective outcome within a campaign. We caution that these individual estimates should not be treated as statistically independent observations because some campaigns reuse the same human ad condition, which creates dependence across the difference estimates.

L.2 Pooled Results Across Studies 1-3 with Study Interactions

Table A24 reports pooled regression results across Studies 1-3 and includes interaction terms that let the CTR and ER effects of AI video ads differ by study.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Overall	TikTok	Meta	Overall	TikTok	Meta
Dep. Var.:	(Pooled)	(Pooled)	(Pooled)	(Pooled)	(Pooled)	(Pooled)
	CTR	CTR	CTR	ER	ER	ER
<i>Video ads</i>						
Human (baseline)						
AI	.1733*** (.0365)	.2577*** (.0409)	-.0867 (.0679)	-.0456* (.0181)	.0179 (.0188)	-.0954*** (.0244)
Study 1 (baseline)						
AI × Study 2	-.0233 (.0662)	-.0643 (.0753)	.2156 (.1580)	.1631*** (.0390)	.1396*** (.0348)	.1059* (.0516)
AI × Study 3	-.0471 (.0782)	-.1220† (.0888)	-.0454 (.1524)	.5658*** (.1366)	.4408** (.1682)	.4292*** (.0997)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	6,178	3,323	2,855	6,178	3,323	2,855
Squared correlation	.0120	.0063	.0109	.1567	.0944	.3590
Pseudo R ²	.9278	.9419	.8852	.9823	.9852	.9844
BIC	13,400	8,230	5,187	81,690	33,595	36,695

Table A24: Regression results with study interaction, Studies 1-3

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination. Based on data from Studies 1-3.

L.3 Winsorization

We obtain consistent results when we winsorize CTR and ER at the 1st and 99th percentiles for the pooled results, see Table A25.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Overall	TikTok	Meta	Overall	TikTok	Meta
Dep. Var.:	(Pooled)	(Pooled)	(Pooled)	(Pooled)	(Pooled)	(Pooled)
	CTR	CTR	CTR	ER	ER	ER
<i>Video ads</i>						
Human (baseline)						
AI	.1804*** (.0285)	.2358*** (.0311)	-.0457 (.0514)	.0481 (.0323)	.0912** (.0323)	-.0329 (.0278)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	2,646	1,612	1,034	5,208	2,523	2,685
Squared correlation	.4251	.2467	.6049	.4604	.4697	.5349
Pseudo R ²	.9413	.9503	.9169	.9820	.9853	.9842
BIC	10,641	5,944	3,664	82,917	33,358	36,981

Table A25: Regression winsorization results, Study 1-3

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination. Based on data from Studies 1-3. Variables are winsorized at the 1st and 99th percentiles. Because multiple observations may share the same value at the percentile cutoffs, the proportion of observations affected can exceed 1% in either tail.

M Lab Studies: Replication and Heterogeneous Effects

Setup

Each lab study begins with participants completing an Informed Consent Agreement followed by a Captcha verification. We then administer two attention checks that simultaneously assess English language proficiency. Since the video ads contain audio elements, we implement an additional audio capability check, excluding participants who fail this verification. Following the attention checks, the respondents watch a 10-second video ad and respond to multiple items as detailed in Table A26. To minimize response bias related to preconceptions about AI-generated content, we deliberately avoid mentioning any connection to AI until the final question.

Construct (reliability)	Source	Items (measured on 7-point Likert scales)
Click intention	Yoo (2007)	<i>How likely is it that you will click on the video ad?</i> (anchored by not at all likely/extremely likely)
Stream intention	Yoo (2007)	<i>How likely would you stream the song?</i> Anchored by not at all likely/extremely likely
Engagement ($\alpha = .94$)	Cian et al. (2014)	<i>Rate the extend to which you think the video is...</i> i. engaging ii. involving iii. boring (R) iv. stimulating (all anchored by not at all/extremely)
Perceived artificiality	Jakesch et al. (2023)	<i>Is the video ad human-made, OR AI-generated?</i> (anchored by definitely human-made/definitely AI-generated)
Quality	Zhang et al. (2022)	<i>Give a score to the video ad on a scale of 1-7 on its aesthetic quality where 1 is very bad and 7 is excellent.</i> (anchored by very bad/excellent)
Creativity ($\alpha = .72$)	Smith et al. (2007)	<i>All things considered, how creative is the video ad compared to an average video ad?</i> i. In general, the video ad was very creative. ii. The video ad should win an award for creativity. iii. The video ad was not very inventive and displayed little creativity in its design. (R) (all anchored by strongly disagree/strongly agree)

Table A26: Overview of scales used for the lab studies

Notes: (R) indicates reverse-coded items.

Lab Study 1: Results

This gender-balanced, controlled study includes 500 respondents (49% women, average age: 40.4 years, UK based), whom we assigned randomly to 50 conditions (25 human and 25 AI) in a between-subjects design on Prolific. From our initial 500 collected responses, eight participants were excluded (six failing the first attention check, one failing the second attention check, and one failing the audio check), yielding 492 valid responses for analysis. The human assets are the human video ads from Study 1, with the objective of a “near-exact replication” (Boegershausen et al., 2025). The AI assets are randomly sampled from the AI video ads from Study 1, with one video ad per campaign.

The AI video ads scored significantly higher on click intentions ($M_{AI} = 2.96$, $SD = 1.86$; $M_{Human} = 2.53$, $SD = 1.66$; $t(476) = 2.74$, $p = .006$) and on engagement ($M_{AI} = 3.25$, $SD = 1.53$, $M_{Human} = 2.80$, $SD = 1.43$; $t(482) = 3.39$, $p < .001$; Cronbach’s $\alpha = .89$) compared to the human video ads.

Table A27 presents the regression results including music track fixed effects. Consistent with the field data, we find a significant positive effect of AI video ads on click intention ($\beta_{AI} = .4555$, $p = .0052$) and engagement ($\beta_{AI} = .4582$, $p < .001$). Notably, there is no detectable difference in terms of perceived artificiality ($\beta_{AI} = -.0244$, $p = .8509$).

Model:	(1)	(2)	(3)	(4)	(5)	(6)
DVs:	Click intention	Stream intention	Engagement	Quality	Creativity	Perceived artificiality
<i>Video ads</i>						
Human (baseline)						
AI	.4555** (.1624)	.3489† (.1847)	.4582*** (.1374)	.6627*** (.1423)	.4450*** (.1266)	-.0244 (.1296)
<i>Fixed Effects:</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fit statistics</i>						
Observations	492	492	492	492	492	492
R ²	.0854	.0901	.0596	.1006	.1016	.1075
Within R ²	.0170	.0077	.0235	.0434	.0268	.0001

Table A27: Regression results, Lab Study 1

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

Lab Study 2: Results

This gender-balanced, controlled study includes 1,207 respondents (49% women, average age: 42.2 years, US based), whom we assigned randomly to 12 conditions (six human and

six AI) in a between-subjects design on CloudResearch. From our initial 1,207 collected responses, 22 participants were excluded (eleven failing the first attention check, six failing the second attention check, and five failing the audio check), yielding 1,185 valid responses for analysis. The assets originate from six campaigns from the first lab study. We again obtain ratings for the same dimensions as in Lab Study 1. Lab Study 2 is preregistered (AsPredicted #242528).

This study again confirms that AI video ads are more clickable ($\beta = .1766$, $p = .0918$) and more engaging ($\beta = .3742$, $p < .001$), see Table A28, while differences in perceived artificiality are again insignificant ($\beta = .0989$, $p = .2283$). The larger number of observations per ad allows us to explore heterogeneous effects of AI video ads across demographics, which we report in Table A29.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
DVs:	Click intention	Stream intention	Engagement	Quality	Creativity	Perceived artificiality
<i>Video ads</i>						
Human (baseline)						
AI video ads	.1766† (.1047)	.0532 (.1109)	.3742*** (.0906)	.2813** (.0864)	.4012*** (.0829)	.0989 (.0821)
<i>Fixed Effects:</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fit statistics</i>						
Observations	1,185	1,185	1,185	1,185	1,185	1,185
R ²	.0123	.0224	.0297	.0578	.0419	.0133
Within R ²	.0024	.0002	.0143	.0089	.0195	.0012

Table A28: Regression results, Lab Study 2

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

N Heterogeneous Effects

N.1 Heterogeneous Effects of AI Video Ads Across Audiences

By pooling observations across Studies 1-3, we find the highest effectiveness in terms of CTR among 18-24-year-olds ($\beta = .2801$, $p < .001$), with net heterogeneous treatment effects (HTE) for AI ($\beta_{AI} + \beta_{AI \times Age}$) by age bin shown in Table A29 for the moderation effects $\beta_{AI \times Age}$. The net effects for CTR suggest a downward gradient of effectiveness, which is supported by a negative interaction between the indexed age bin $A_i \in \{0, 1, 2, 3, 4\}$, corresponding to the age bins 18-24 through 55+, and the AI treatment ($\beta = -.0782$, $p < .001$; plotted in Web Appendix N.3). One possible explanation is that these patterns reflect generational differences in media consumption and AI receptivity (Widener et al., 2025), though our data do not directly test the proposed mechanism, and it therefore remains speculative.

Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Overall	Overall	Lab	Lab	Overall	Overall	Lab	Lab
Dep. Var.:	(Pooled)	(Pooled)	Exp.	Exp.	(Pooled)	(Pooled)	Exp.	Exp.
	CTR	CTR	Click	Click	ER	ER	Engagement	Engagement
	Net	Age	Net	Age	Net	Age	Net	Age
	HTE	Gradient	HTE	Gradient	HTE	Gradient	HTE	Gradient
<i>Video ads</i>								
Human (baseline)								
AI effect, Age: 18-24	.2801*** (.0407)		.7389† (.4101)		.0046 (.0311)		.3318 (.3453)	
AI effect, Age: 25-34	.0822 (.0670)		.2936 (.2144)		-.0118 (.0278)		.5102** (.1752)	
AI effect, Age: 35-44	.1859* (.0864)		.3650† (.1937)		.0445 (.0349)		.5463** (.1660)	
AI effect, Age: 45-54	.0097 (.0999)		-.2360 (.2265)		.0068 (.0336)		.1038 (.2036)	
AI effect, Age: 55+	-.0398 (.0649)		.0969 (.2409)		.1116 (.0772)		.3302 (.2192)	
AI		.2628*** (.0367)		.5355* (.2215)		-.0145 (.0282)		.5522** (.1865)
AI × Age (indexed)		-.0782*** (.0178)		-.1511† (.0880)		.0266 (.0201)		-.0709 (.0766)
<i>Controls</i>								
Audience age	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>								
Music track	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	Yes	No	No	Yes	Yes	No	No
<i>Fit statistics</i>								
Observations	5,617	5,617	1,179	1,179	5,617	5,617	1,179	1,179
Squared correlation	.0096	.0084	.0288	.0275	.1654	.1652	.0423	.0404
Pseudo R ²	.9296	.9296	.0079	.0073	.9821	.9821	.0116	.0110
BIC	13,007	12,985	4,829	4,811	82,344	82,389	4,491	4,473

Table A29: Regression results, pooled across Studies 1-3 and next to lab experiment

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

Models 2, 4, 6, and 8 estimate an age-gradient specification that replaces age-bin indicators with an ordered age index, $A_i \in \{0, 1, 2, 3, 4\}$, corresponding to the age bins 18-24, 25-34, 35-44, 45-54, and 55+, respectively. The reported interaction coefficient captures the linear change in the AI effect across adjacent age bins. Web Appendix N.2 presents a robustness check including a moderation with release year.

A preregistered lab experiment (1,185 valid responses, 49% women, six human and six AI conditions, 99 responses per condition on average, AsPredicted #242528; Web Appendix M) is directionally consistent with this finding: 18-24-year-old respondents show the highest click intention for AI assets ($\beta = .7389$, $p = .0718$) and the downward age gradient is also marginally significant ($\beta = -.1511$, $p = .0860$). For engagement, we observe no significant age gradient in the field or lab (see models 6 and 8).

N.2 Heterogeneous Effects of AI Video Ads Controlling for Release Year

This subsection evaluates whether heterogeneous effects by audience age are robust to controlling for release year. Table A30 reports the corresponding regression results for clicks and engagement pooled across Studies 1-3 and the lab experiment.

Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Overall	Overall	Lab	Lab	Overall	Overall	Lab	Lab
Dep. Var.:	(Pooled)	(Pooled)	Exp.	Exp.	(Pooled)	(Pooled)	Exp.	Exp.
	CTR	CTR	Click Intention	Click Intention	ER	ER	Engagement	Engagement
<i>Video ads</i>								
Human (baseline)								
AI	.1648*** (.0287)	.2798*** (.0406)	.2035† (.1048)	.7309† (.4119)	.0598† (.0313)	.0070 (.0309)	.3929*** (.0905)	.3122 (.3468)
<i>Moderation effects</i>								
AI × Release year _z	.0116 (.0279)	-.0049 (.0281)	.0378 (.1054)	.0350 (.1059)	.0593** (.0183)	.0717*** (.0170)	.0987 (.0899)	.0995 (.0902)
AI × 25-34		-.1976* (.0784)		-.4361 (.4654)		-.0172 (.0415)		.1997 (.3884)
AI × 35-44		-.0944 (.0957)		-.3691 (.4556)		.0448 (.0476)		.2138 (.3852)
AI × 45-54		-.2709* (.1085)		-.9684* (.4701)		.0149 (.0456)		-.2124 (.4018)
AI × 55+		-.3212*** (.0778)		-.6321 (.4788)		.1327 (.0820)		.0217 (.4114)
<i>Controls</i>								
Audience age	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>								
Music track	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	Yes	No	No	Yes	Yes	No	No
<i>Fit statistics</i>								
Observations	6,178	6,178	1,185	1,185	6,178	6,178	1,185	1,185
Squared Cor.	.0120	.0117	.0291	.0342	.1567	.1570	.0406	.0436
Pseudo R ²	.9278	.9280	.0194	.0245	.9818	.9818	.0254	.0285
BIC	13,392	13,405	4,841	4,870	84,963	83,784	4,500	4,532

Table A30: Regression results with release year interaction, pooled across Studies 1-3 and next to lab experiment

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. For parsimony, users with unknown ages are not displayed but included in the regression analyses. Release year_z is standardized ($M_{Release\ year} = 1986.31$, $SD_{Release\ year} = 13.39$). Based on data from Studies 1-3 and lab experiment.

N.3 Age Gradient for CTR Effect of AI Video Ads

Study 3 indicates heterogeneous CTR effects of AI video ads across different age bins (see Table A29). Figure A6 visualizes the CTR effect of AI video ads by age bin in the field and in the lab. The overall pattern suggests a downward age gradient.

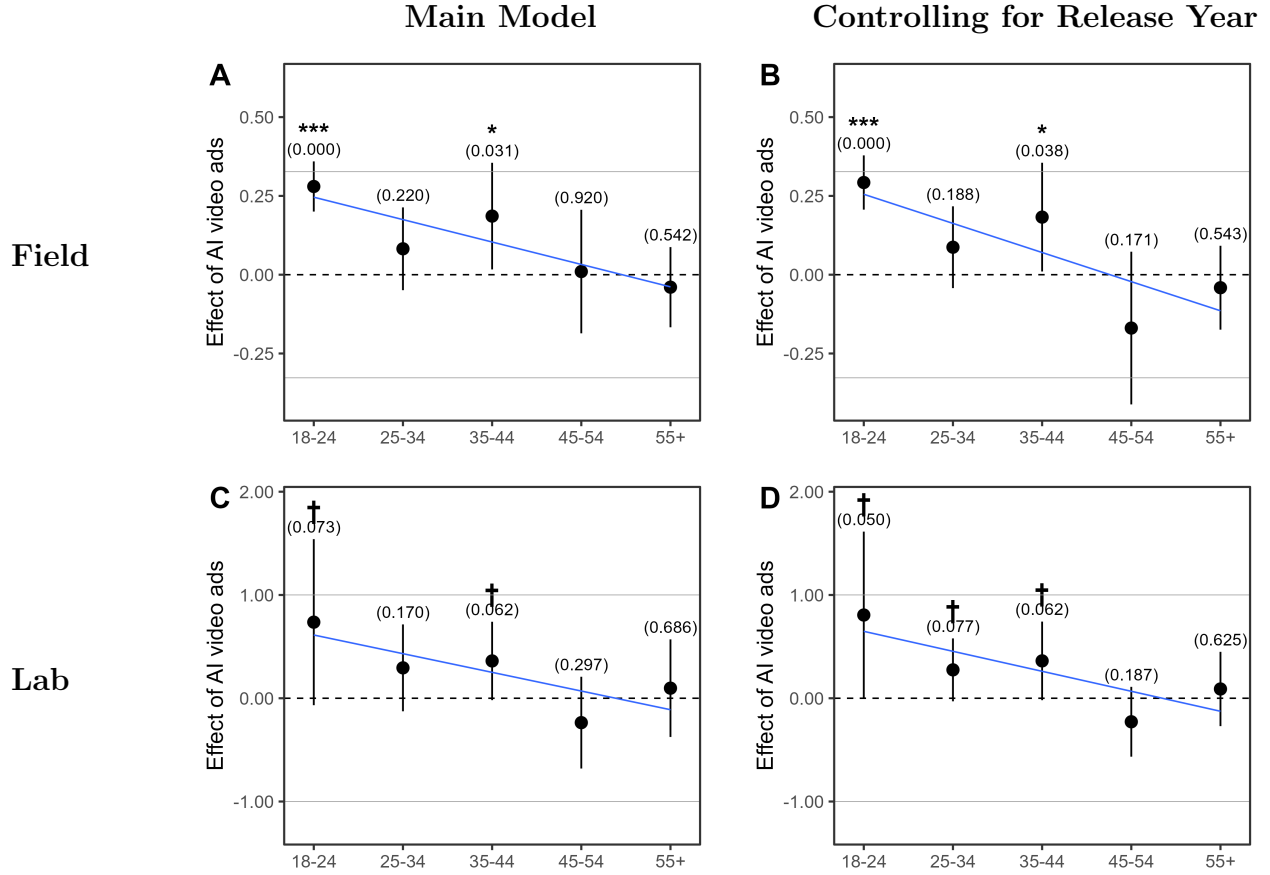


Figure A6: Heterogeneous CTR effects of AI video ads by age bin

Notes: Panels A and C correspond to the coefficients from models (2) and (4) from Table A29. Panels B and D include a control for the release year in models (2) and (4), see Table A30. The blue line represents the fitted linear trend from the point estimates on ordered age bins.

N.4 Analysis of Visual Features and Audience Response

The large number of AI assets across all studies enables exploratory feature-level analysis of how the response of different audiences is associated with specific visual elements, which can be circulated back to our industry partner for learning. Because features are not independently randomized, these results are exploratory and descriptive rather than causal. For example, the presence of an intro shot is associated with increased engagement among

audiences older than 55 on TikTok (male: $\beta = .4657$, $p < .001$; female: $\beta = .4765$, $p < .001$) and Meta (male: $\beta = .4649$, $p < .001$; female: $\beta = .5046$, $p < .001$), but negatively correlated with engagement among younger audiences. These effects remain consistent when accounting for multiple hypothesis testing with Bonferroni adjustments. Tables A31 and A32 present complete feature-level effects of AI video ads across audience segments for CTR and ER, respectively.

Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Platform:	TikTok	TikTok	TikTok	TikTok	Meta	Meta	Meta	Meta
Gender:	male	female	male	female	male	female	male	female
Age:	18-24	18-24	55+	55+	18-24	18-24	55+	55+
Dep. Var.:	CTR	CTR	CTR	CTR	CTR	CTR	CTR	CTR
<i>Narrative:</i>								
Intro	.4887*** (.1017)	.6766*** (.0810)	.2688† (.1517)	.6829** (.2083)	-.1919 (.2868)	-.0906 (.4623)	-.6638* (.2658)	-1.638* (.7150)
Outro	-.1854 (.1493)	.0648 (.2023)	.1698 (.3604)	-.3940 (.3902)	-.0330 (.2822)	-.2841 (.6179)	-.2761 (.3836)	.3842 (.9179)
<i>Content:</i>								
Position: Middle	-.0123 (.1170)	-.0360 (.1295)	.1637 (.1504)	.0482 (.1553)	.2687 (.1735)	-.1692 (.3479)	-.3292 (.2187)	.1913 (.4197)
Position: Top	-.0230 (.1067)	-.1111 (.1199)	.1536 (.1997)	.0394 (.1809)	.0323 (.2967)	.4551 (.5123)	-.0547 (.2462)	.7344 (.5648)
Size: Medium	.1657† (.0888)	-.0071 (.0826)	.0995 (.1521)	-.0165 (.1582)	-.0770 (.2669)	.3976 (.4249)	-.1073 (.1693)	.7345† (.4176)
Size: Small	-.0529 (.1242)	-.0560 (.1292)	-.0070 (.2022)	-.0353 (.2042)	-.1808 (.2133)	.1343 (.4036)	.2305 (.1911)	.3076 (.5061)
Edge radius: Small	.0073 (.0927)	-.0016 (.0848)	-.2734* (.1195)	.0366 (.1425)	.1108 (.2488)	.7934* (.3264)	.0700 (.1612)	.2777 (.3869)
Edge radius: Large	.0801 (.0771)	.0779 (.0891)	-.0639 (.1563)	-.0961 (.1622)	-.0626 (.2225)	.0325 (.3738)	.0515 (.1516)	.5468 (.4759)
Shadow	-.0175 (.1006)	.1016 (.0809)	-.0634 (.1528)	-.3224* (.1549)	-.3612 (.2814)	-.4395 (.3766)	.1227 (.1499)	.0806 (.4193)
Waveform	-.0692 (.0706)	-.1315† (.0755)	.0361 (.1201)	-.1796 (.1155)	.0352 (.2301)	-.0566 (.2824)	.0061 (.1159)	.0887 (.3555)
<i>Animation:</i>								
Shiny	.1835 (.1155)	.2165† (.1122)	-.1012 (.1776)	.2742 (.2560)	.6287 (.4488)	-.8453 (.5484)	-.2515 (.2639)	.2081 (.5500)
Product	.1207 (.3629)	.2911 (.1983)	-16.41*** (.4597)	.0968 (.5577)	-16.30*** (1.295)	-14.62*** (.7664)	1.523 (1.025)	.6997 (1.314)
Reflection	-.1905† (.1119)	.0341 (.1065)	-.3113† (.1756)	-.4554** (.1634)	-.1173 (.2953)	-.2962 (.4343)	.4620* (.2205)	-.4838 (.4343)
Fly-in	-.2932* (.1490)	.0578 (.1310)	-.5473* (.2140)	-.0858 (.2918)	-.0966 (.3222)	-.4310 (.4642)	-.0808 (.2322)	.7099 (.5580)
Zoom	-.0254 (.1220)	.0415 (.1215)	.0483 (.1786)	.0941 (.1892)	.0093 (.2770)	-.4573 (.4811)	-.0718 (.2277)	-.0899 (.5129)
Boom	-.0881 (.2026)	-.1722 (.1601)	-.4293 (.3288)	-.4973 (.4154)	-.1604 (.4767)	-1.324† (.6840)	-.2776 (.4567)	.2193 (.8533)
Shear	.2915 (.3215)	-.5077* (.2159)	-17.39*** (.9406)	-.6059† (.3337)	.5195 (.5034)	-13.65*** (1.232)	.6308 (.4757)	1.784 (1.533)
<i>Background:</i>								
Color	.8276* (.3942)	-.0703 (.4141)	.7594 (.6208)	-.0390 (.7770)	1.015 (1.048)	-.2544 (1.488)	-.2268 (.6288)	-2.684 (1.950)
Packshot	.0746 (.1741)	.0972 (.1781)	-.2860 (.3111)	.3935 (.3694)	.7664 (.4714)	-.4521 (.6753)	-.3377 (.4657)	-2.654* (1.084)
GenAI Image	.2139 (.1831)	.0970 (.2164)	.0733 (.3415)	.2086 (.4273)	.4528 (.5251)	.0567 (.7541)	-.6464 (.5534)	-2.332† (1.362)
GenAI Images	.6013† (.3479)	.4876 (.4252)	-.0746 (.5597)	.5193 (.6221)	.4848 (.8055)	-2.397† (1.228)	-1.237† (.7390)	-3.204† (1.732)
GenAI Videos	.5303 (.3516)	.4874 (.3944)	.1021 (.5416)	.3626 (.5886)	-.1580 (.8002)	-1.676 (1.118)	-1.000 (.6895)	-2.683 (1.685)
GenAI 2× Videos	.7253† (.3707)	.6970 (.4510)	.0849 (.6530)	.7465 (.7071)	-.0678 (.9450)	-3.300* (1.325)	-1.604* (.7799)	-2.555 (1.839)
Animation: Move	.2657 (.1639)	.1613 (.1655)	.2152 (.2482)	-.2735 (.2865)	-.4247 (.3415)	-.5923 (.4861)	-.3163 (.3089)	-.8636 (.7701)
Animation: Multiple	-.2895 (.2085)	-.3531 (.2643)	.0678 (.2605)	-.4981* (.2510)	-.2949 (.3784)	.9531 (.5827)	-.1883 (.3051)	-.1706 (.6216)
Motion Bucket ID	.0001 (.0005)	.00002 (.0005)	.0002 (.0008)	-.0002 (.0008)	.0029† (.0016)	.0017 (.0023)	.0006 (.0010)	.0001 (.0021)
Brightness: Light	-.2625** (.0871)	-.2119* (.0843)	-.2588* (.1259)	-.0057 (.1910)	.2057 (.2703)	.1195 (.3661)	.1595 (.1555)	-.2142 (.3357)
<i>Text:</i>								
Position: Below	.1117 (.0749)	.0382 (.0826)	-.0748 (.1207)	.0981 (.1326)	-.0009 (.2142)	.2942 (.3002)	-.1852 (.1443)	.0239 (.3988)
Font: Medium	.0156 (.0791)	.0558 (.0782)	-.4125*** (.1221)	-.0921 (.1466)	-.4830* (.2136)	-.1472 (.3238)	-.1525 (.1529)	.2750 (.4940)
Font: Small	-.0443 (.0906)	-.0448 (.0930)	-.2255 (.1536)	-.0425 (.1382)	-.2319 (.2156)	-.6692* (.3018)	-.0853 (.1323)	.4484 (.4465)
Color: Custom	-.1116 (.0836)	-.0121 (.0783)	.2029† (.1167)	.1405 (.1357)	-.1216 (.2263)	-.3728 (.3489)	-.0178 (.2009)	.8421† (.4686)
Color: White	-.0159 (.1021)	.0126 (.1049)	-.0191 (.1421)	.2547 (.1718)	.3360 (.2979)	.0763 (.4272)	.3017 (.2150)	1.278** (.4407)
Animation	-.2670 (.2219)	-.1078 (.2183)	-.6365* (.3079)	-.2690 (.5068)	-16.83*** (.3986)	1.305† (.6745)	-.1687 (.3642)	-.7945 (1.041)
<i>Audio:</i>								
Louder	.0106 (.0683)	-.0163 (.0726)	-.0354 (.1054)	-.0830 (.1167)	-.2498 (.1784)	-.0137 (.3171)	-.2866† (.1654)	.1155 (.2898)
Speed: Slow @90%	.0315 (.1080)	.2179* (.1002)	.0091 (.1744)	-.1221 (.2030)	.1828 (.3487)	.0921 (.4511)	-.3788† (.2000)	-.2685 (.4204)
Speed: Fast @105%	-.0628 (.1346)	.1708 (.1282)	-.0985 (.2980)	-.2386 (.3404)	-.4218 (.2723)	-.5784 (.4795)	-.1086 (.3063)	1.077 (.7879)
Speed: Fast @110%	.0326 (.1103)	.0763 (.1058)	-.0500 (.1580)	.0556 (.1834)	-.1627 (.3211)	1.020* (.4445)	-.2298 (.2144)	-.2500 (.4819)
<i>Fixed effects</i>								
Music track	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fit statistics</i>								
Observations	164	164	164	164	160	144	164	136
Squared Cor.	.6764	.6860	.4521	.3477	.2838	.3314	.4736	.7218
Pseudo R ²	.9716	.9764	.9464	.9382	.9171	.8977	.9429	.9021
BIC	1,120	1,203	830	845	698	557	805	564

Table A31: Regression results for CTR per audience segment based on video features
Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.
Based on data from Studies 1-3.

Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Platform:	TikTok	TikTok	TikTok	TikTok	Meta	Meta	Meta	Meta
Gender:	male	female	male	female	male	female	male	female
Age:	18-24	18-24	55+	55+	18-24	18-24	55+	55+
Dep. Var.:	ER	ER	ER	ER	ER	ER	ER	ER
<i>Narrative:</i>								
Intro	-.2318*** (.0572)	-.1857*** (.0488)	.4657*** (.0762)	.4765*** (.0703)	-.0997† (.0586)	-.0726 (.0696)	.4649*** (.0603)	.5046*** (.0600)
Outro	.1929 (.1353)	.0829 (.0959)	.2191* (.0919)	.0879 (.1217)	-.0087 (.0957)	.1933† (.1065)	-.0344 (.0625)	-.1607 (.1029)
<i>Content:</i>								
Position: Middle	-.0402 (.0463)	-.0065 (.0472)	.0226 (.0530)	.0135 (.0512)	.0065 (.0408)	-.0020 (.0585)	-.0615 (.0528)	-.1407* (.0634)
Position: Top	-.0365 (.0442)	.0012 (.0461)	.0613 (.0579)	-.0227 (.0597)	.0335 (.0557)	.0197 (.0752)	-.0630 (.0626)	-.1528* (.0757)
Size: Medium	-.0180 (.0407)	.0176 (.0338)	-.0173 (.0604)	.0266 (.0550)	-.0381 (.0444)	.0166 (.0558)	-.0262 (.0478)	-.0182 (.0503)
Size: Small	-.0920† (.0499)	-.0249 (.0493)	-.0673 (.0696)	-.0651 (.0687)	-.0871† (.0506)	-.0896 (.0702)	-.0714 (.0666)	.0367 (.0735)
Edge radius: Small	.0079 (.0327)	-.0165 (.0308)	-.0304 (.0485)	-.0265 (.0462)	-.0493 (.0416)	-.0269 (.0551)	-.0115 (.0445)	-.1320*** (.0439)
Edge radius: Large	-.0187 (.0315)	-.0125 (.0350)	-.0495 (.0552)	-.0579 (.0581)	-.0708† (.0427)	-.0893† (.0499)	-.0031 (.0457)	-.0768 (.0510)
Shadow	-.0302 (.0426)	-.0042 (.0404)	-.0140 (.0663)	.0317 (.0585)	.0547 (.0466)	.0269 (.0573)	.0138 (.0515)	.0899 (.0607)
Waveform	.0492 (.0307)	.0155 (.0317)	.0587 (.0376)	-.0282 (.0372)	.0796* (.0331)	-.0057 (.0532)	.0206 (.0384)	.0013 (.0447)
<i>Animation:</i>								
Shiny	.0082 (.0529)	.0159 (.0461)	.0169 (.0721)	.0138 (.0803)	-.1530* (.0620)	.0391 (.0795)	-.0242 (.0621)	.0286 (.0740)
Product	-.0516 (.2053)	-.2239** (.0704)	-.4479* (.1891)	-.1042 (.1391)	.0784 (.2174)	-.0731 (.2309)	-.2502 (.1962)	-.1889 (.1194)
Reflection	.0429 (.0430)	.0025 (.0449)	.1241* (.0583)	.1258* (.0510)	-.0681 (.0466)	-.0146 (.0639)	.0199 (.0522)	.1749** (.0635)
Fly-in	.0508 (.0620)	-.0326 (.0552)	.1895** (.0639)	.2690*** (.0607)	-.1498* (.0717)	.0952 (.1019)	.0185 (.0676)	.3243*** (.0725)
Zoom	.0030 (.0529)	.0041 (.0471)	-.0512 (.0628)	-.0071 (.0632)	-.0029 (.0565)	.0457 (.0687)	-.0939 (.0617)	.0072 (.0596)
Boom	.0247 (.0927)	-.0069 (.0863)	.0151 (.1112)	.1467 (.1119)	.1238 (.0971)	-.0392 (.1236)	.0338 (.0988)	.0824 (.1038)
Shear	.4003*** (.0836)	.3890** (.1244)	.1516 (.1220)	.3016* (.1242)	.2168 (.1456)	.2437* (.1183)	-.0599 (.1484)	-.0634 (.1284)
<i>Background:</i>								
Color	.1542 (.1371)	.1462 (.1685)	.0157 (.2208)	.0146 (.2034)	.1314 (.1960)	-.2379 (.2381)	-.1244 (.1844)	-.0501 (.2463)
Packshot	-.0208 (.0806)	.1350 (.0828)	-.0378 (.1361)	.0714 (.1476)	.0274 (.1072)	-.1916† (.1126)	-.0138 (.0819)	.0255 (.0955)
GenAI Image	.0406 (.0823)	.1188 (.0861)	.0438 (.1539)	.0290 (.1683)	-.0527 (.1193)	-.2656* (.1151)	-.1307 (.1116)	-.1216 (.1305)
GenAI Images	.0351 (.1255)	.2412 (.1571)	-.0718 (.2151)	.0061 (.2470)	.0426 (.1890)	-.4550* (.1910)	-.0615 (.1884)	-.0384 (.1805)
GenAI Videos	.0201 (.1226)	.2765† (.1435)	.0862 (.2167)	.1020 (.2221)	.1065 (.1843)	-.3852* (.1722)	.0184 (.1834)	.0337 (.1761)
GenAI 2× Videos	-.0469 (.1303)	.2237 (.1522)	-.0526 (.2538)	-.0094 (.2617)	.1896 (.2012)	-.3698† (.1910)	-.0413 (.2037)	-.1024 (.1973)
Animation: Move	.1400* (.0692)	.2072** (.0660)	.0892 (.0714)	.0894 (.0754)	.1076 (.0772)	-.0259 (.0839)	.0051 (.0727)	-.0528 (.0843)
Animation: Multiple	.0691 (.0603)	.0341 (.0858)	.1296 (.1185)	.0858 (.1056)	-.0077 (.0919)	.1942* (.0865)	.0327 (.1174)	-.0042 (.0886)
Motion Bucket ID	.0002 (.0002)	-.0001 (.0002)	-.0001 (.0002)	.000005 (.0003)	-.0001 (.0002)	.0002 (.0003)	-.0003 (.0002)	-.0006* (.0003)
Brightness: Light	-.0163 (.0373)	-.0475 (.0339)	-.1747*** (.0493)	-.1347** (.0485)	-.0988* (.0401)	.0439 (.0462)	-.1044* (.0453)	-.1460** (.0478)
<i>Text:</i>								
Position: Below	-.0417 (.0305)	-.0554† (.0321)	.0679 (.0455)	.0387 (.0396)	.0219 (.0396)	-.0093 (.0519)	.0228 (.0474)	.0038 (.0547)
Font: Medium	.0329 (.0303)	.0245 (.0318)	.0136 (.0425)	.0602 (.0478)	-.0447 (.0351)	.0343 (.0511)	.0250 (.0360)	.0399 (.0528)
Font: Small	-.0068 (.0370)	.0392 (.0358)	-.0938† (.0517)	-.0133 (.0459)	-.0387 (.0451)	.0585 (.0509)	.0074 (.0435)	.0308 (.0513)
Color: Custom	-.1024** (.0343)	-.0548† (.0311)	-.0895* (.0352)	-.0844* (.0342)	-.0495 (.0382)	-.0325 (.0540)	.0179 (.0384)	-.0465 (.0467)
Color: White	-.0504 (.0436)	-.0347 (.0406)	.0293 (.0508)	.0252 (.0551)	-.0964* (.0485)	.0892 (.0776)	-.0263 (.0524)	-.0299 (.0657)
Animation	.0220 (.0669)	-.0633 (.0634)	.1277 (.0950)	.1062 (.1074)	-.1088 (.1077)	.0829 (.1486)	.0458 (.0911)	-.1110 (.1230)
<i>Audio:</i>								
Louder	.0137 (.0253)	.0480* (.0241)	-.0201 (.0367)	-.0192 (.0386)	.0123 (.0315)	-.0529 (.0399)	-.0517 (.0333)	-.0821* (.0329)
Speed: Slow @90%	-.0238 (.0428)	.0164 (.0426)	.0012 (.0711)	.0015 (.0668)	.0764 (.0567)	-.0496 (.0662)	-.0653 (.0625)	-.0889 (.0672)
Speed: Fast @105%	.0190 (.0720)	-.0482 (.0571)	-.2093** (.0707)	-.2852*** (.0806)	.0234 (.0558)	.0139 (.0699)	-.0891† (.0513)	-.0192 (.0576)
Speed: Fast @110%	.0137 (.0444)	-.0067 (.0407)	.0845 (.0558)	.0454 (.0508)	.0789† (.0442)	.0005 (.0585)	.0202 (.0543)	.0300 (.0574)
<i>Fixed effects</i>								
Music track	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fit statistics</i>								
Observations	164	164	164	164	164	164	164	164
Squared Cor.	.8302	.8149	.8558	.8892	.5956	.4746	.83712	.8355
Pseudo R ²	.9946	.9954	.9945	.9941	.9902	.9820	.9958	.9935
BIC	1,807	1,817	1,980	1,849	1,466	1,268	2,543	1,704

Table A32: Regression results for ER per audience segment based on video features
Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.
Based on data from Studies 1-3.

O Robustness Analyses

We conduct several supplementary analyses with consistent results if we (1) divide clicks and engagement by reach (cf. impressions; Web Appendix O.1), (2) aggregate across age and gender (Web Appendix O.2)⁶, (3) estimate OLS regressions (cf. binomial regressions; Web Appendix O.3), (4) winsorize CTR and ER at the 1st and 99th percentiles (Web Appendix O.4), and (5) control for each video ad’s average pairwise visual similarity to all other video ads within the same campaign, using 768-dimensional, real-valued vector embeddings from a video vision transformer (ViViT (Arnab et al., 2021); Web Appendix P).

O.1 Reach

As a robustness check, we also compute the performance metrics for clicks and engagement divided by reach rather than impressions (Boegershausen et al., 2025). This alternative calculation yields consistent results across Studies 1-3, as demonstrated in Tables A33, A34, and A35.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall	TikTok	Meta	Overall	TikTok	Meta
	CTR	CTR	CTR	ER	ER	ER
	by reach	by reach	by reach	by reach	by reach	by reach
<i>Video ads</i>						
Human (baseline)						
AI	.1573*** (.0366)	.2364*** (.0418)	-.0825 (.0675)	-.0534** (.0187)	-.0013 (.0193)	-.0970*** (.0248)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	3,125	1,250	1,875	3,124	1,250	1,874
Squared correlation	.0265	.1782	.0021	.5132	.7446	.4905
Pseudo R ²	.9264	.9428	.8797	.9840	.9167	.9841
BIC	9,421	5,499	3,934	47,657	341	24,212

Table A33: Regression results for based on reach, Study 1

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

The unit of observation is an ad-platform-age-gender combination.

⁶Recall Study 3 is narrowly targeted to the most granular gender and age bins that we observe. Hence, unlike in Study 1 and 2, no aggregation is possible.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall CTR by reach	TikTok CTR by reach	Meta CTR by reach	Overall ER by reach	TikTok ER by reach	Meta ER by reach
<i>Video ads</i>						
Human (baseline)						
AI: randomly sampled (RS)	.1614** (.0620)	.1733* (.0694)	.2847† (.1557)	.0871** (.0301)	.1334*** (.0324)	.0383 (.0461)
AI: expert selected (ES)	.0702 (.0667)	.1331† (.0759)	.0215 (.1533)	.0501 (.0306)	.1005** (.0361)	-.0007 (.0454)
Difference between AI conditions (RS – ES)	.0912 (.0621)	.0402 (.0671)	.2633* (.1208)	.0370 (.0288)	.0329 (.0287)	.0390 (.0418)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	1,574	630	944	1,573	630	943
Squared correlation	.0178	.0383	.0109	.3202	.7096	.2385
Pseudo R ²	.9065	.9235	.8466	.9801	.9833	.9782
BIC	3,035	1,960	1,083	17,860	6,908	10,220

Table A34: Regression results for based on reach, Study 2

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

The unit of observation is an ad-platform-age-gender combination.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall CTR by reach	TikTok CTR by reach	Meta CTR by reach	Overall ER by reach	TikTok ER by reach	Meta ER by reach
<i>Video ads</i>						
Human (baseline)						
AI: predicted best CTR (PB)	.1182† (.0699)	.1993*** (.0549)	-.1424 (.1713)	.6553*** (.1207)	.7772*** (.1749)	.2690† (.1440)
AI: predicted worst CTR (PW)	-.0768 (.0759)	.0031 (.0589)	-.3026† (.1802)	.2421† (.1315)	.0125 (.1759)	.4341*** (.1306)
Difference between AI conditions (PB – PW)	.1950** (.0603)	.1962** (.0598)	.1602 (.2229)	.4132*** (.1165)	.7647*** (.0646)	-.1651 (.1600)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	72	36	36	72	36	36
Squared correlation	.8838	.4237	.2647	.8033	.9754	.7739
Pseudo R ²	.9809	.9860	.9619	.9855	.9933	.9894
BIC	393.5	225.8	168.2	7,519	1,978	2,378

Table A35: Regression results based on reach, Study 3

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

The unit of observation is an ad-platform-age-gender combination.

O.2 Aggregated Demographics

When examining results with demographics (age and gender) aggregated, we observe consistent effects in Studies 1 and 2, as detailed in Tables A36 and A37. Study 3 is already targeted to specific demographics and therefore cannot be aggregated any further.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall	TikTok	Meta	Overall	TikTok	Meta
	CTR	CTR	CTR	ER	ER	ER
<i>Video ads</i>						
Human (baseline)						
AI	.1581*** (.0461)	.2369*** (.0543)	-.0867 (.0679)	-.0171 (.0348)	.0761* (.0346)	-.0942* (.0406)
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	250	125	125	250	125	125
Squared correlation	.7246	.4748	.5508	.6946	.8502	.76851
Pseudo R ²	.9851	.98874	.9785	.9943	.9973	.9968
BIC	2,067	1,198	805.2	18,160	4,218	5,368

Table A36: Regression results with demographics aggregated, Study 1

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform combination.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall	TikTok	Meta	Overall	TikTok	Meta
	CTR	CTR	CTR	ER	ER	ER
<i>Video ads</i>						
Human (baseline)						
AI: randomly sampled (RS)	.1941*** (.0519)	.2179*** (.0585)	.2917* (.1449)	.1194** (.0450)	.2163*** (.0395)	.0121 (.0551)
AI: expert selected (ES)	.0859 (.0594)	.1491* (.0719)	.0372 (.1418)	.0643 (.0508)	.1792*** (.0478)	-.0716 (.0536)
Difference between AI conditions (RS – ES)	.1082 (.0689)	.0688 (.0704)	.2545* (.1147)	.0551 (.0499)	.0372 (.0609)	.0837 (.0680)
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	126	63	63	126	63	63
Squared correlation	.6848	.2518	.3624	.7862	.5637	.6295
Pseudo R ²	.9784	.9847	.9628	.9943	.9953	.9944
BIC	732.9	419.0	276.6	5,250	1,995	2,703

Table A37: Regression results with demographics aggregated, Study 2

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform combination.

O.3 OLS

Estimating OLS regressions with CTR and ER (both in %) as the dependent variables also reveals consistent results, see Tables A38, A39, and A40.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall CTR (in %)	TikTok CTR (in %)	Meta CTR (in %)	Overall ER (in %)	TikTok ER (in %)	Meta ER (in %)
<i>Video ads</i>						
Human (baseline)						
AI	.0172*** (.0036)	.0354*** (.0055)	-.0047 (.0038)	-.2043* (.0924)	.0940 (.0821)	-.5592*** (.1482)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	4,114	2,239	1,875	4,114	2,239	1,875
R ²	.3217	.1964	.1058	.76478	.78802	.7893
Within R ²	.06962	.0997	.0356	.6952	.7192	.7448

Table A38: Results with OLS regressions, Study 1

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall CTR (in %)	TikTok CTR (in %)	Meta CTR (in %)	Overall ER (in %)	TikTok ER (in %)	Meta ER (in %)
<i>Video ads</i>						
Human (baseline)						
AI: randomly sampled (RS)	.0202*** (.0059)	.0268** (.0081)	.0128† (.0071)	.5225** (.1931)	.5540*** (.1522)	.3088 (.3366)
AI: expert selected (ES)	.0090 (.0058)	.0184* (.0087)	.0014 (.0065)	.2603 (.1924)	.4241** (.1627)	.0720 (.3050)
Difference between AI conditions (RS – ES)	.0112† (.0057)	.0084 (.0084)	.0114* (.0053)	.2622 (.1719)	.1299 (.1262)	.2368 (.3139)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	
Audience gender	Yes	Yes	Yes	Yes	Yes	
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	1,992	1,048	944	1,992	1,048	944
R ²	.1619	.0540	.0818	.7348	.6874	.6846
Within R ²	.0339	.0427	.0191	.5642	.6445	.5701

Table A39: Results with OLS regressions, Study 2

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall CTR (in %)	TikTok CTR (in %)	Meta CTR (in %)	Overall ER (in %)	TikTok ER (in %)	Meta ER (in %)
<i>Video ads</i>						
AI: predicted best CTR (PB)	.0198* (.0091)	.0430*** (.0113)	-.0039 (.0089)	3.686** (1.138)	4.987** (1.797)	1.351† (.7545)
AI: predicted worst CTR (PW)	-.0047 (.0086)	.0056 (.0106)	-.0127 (.0079)	.6459 (1.010)	-.3781 (1.639)	2.032* (.7973)
Difference between AI conditions (PB – PW)	.0245** (.0085)	.0374** (.0114)	.0088 (.0099)	.3.040** (.9508)	5.365*** (1.227)	-.6812 (.8810)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Platform	Yes	No	No	No	No	
<i>Fit statistics</i>						
Observations	72	36	36	72	36	36
R ²	.8759	.4660	.2062	.7302	.8122	.8150
Within R ²	.2406	–	–	.7249	–	–

Table A40: Results with OLS regressions, Study 3

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination.

O.4 Winsorization

When we winsorize CTR and ER at the 1st and 99th percentiles (calculated separately for each study and platform) we replace observations that exceed the top (bottom) 1% with the corresponding percentile value. We obtain consistent results with winsorization, see Tables A41, A42, and A43.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall	TikTok	Meta	Overall	TikTok	Meta
	CTR	CTR	CTR	ER	ER	ER
<i>Video ads</i>						
Human (baseline)						
AI	.1734*** (.0364)	.2569*** (.0410)	-.0846 (.0680)	-.0445* (.0176)	.0182 (.0186)	-.0956*** (.0237)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	4,114	2,239	1,875	4,114	2,239	1,875
Squared correlation	.0353	.0128	.0604	.2702	.1558	.6119
Pseudo R ²	.9269	.9415	.8849	.9851	.9876	.9862
BIC	9,764	5,826	3,869	47,155	19,026	22,578

Table A41: Regression winsorization results, Study 1

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall CTR	TikTok CTR	Meta CTR	Overall ER	TikTok ER	Meta ER
<i>Video ads</i>						
Human (baseline)						
AI: randomly sampled (RS)	.1962** (.0621)	.2173** (.0694)	.2748† (.1562)	.1013*** (.0297)	.1734*** (.0309)	.0332 (.0473)
AI: expert selected (ES)	.0949 (.0671)	.1579* (.0758)	.0177 (.1538)	.0714* (.0304)	.1380*** (.0346)	-.0043 (.0469)
Difference between AI conditions (RS – ES)	.1013 (.0617)	.0594 (.0667)	.2571* (.1202)	.0298 (.0273)	.0354 (.0282)	.0375 (.0397)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	1,992	1,048	944	1,992	1,048	944
Squared correlation	.0523	.0011	.0506	.2790	.1770	.3105
Pseudo R ²	.9062	.9210	.8526	.9803	.9819	.9797
BIC	3,092	2,071	1,041	17,980	7,693	9,707

Table A42: Regression winsorization results, Study 2

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.
The unit of observation is an ad-platform-age-gender combination.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall CTR	TikTok CTR	Meta CTR	Overall ER	TikTok ER	Meta ER
<i>Video ads</i>						
Human (baseline)						
AI: predicted best CTR (PB)	.1528* (.0720)	.2295*** (.0560)	-.0819 (.1822)	.6712*** (.1111)	.7907*** (.1666)	.3099* (.1344)
AI: predicted worst CTR (PW)	-.0509 (.0779)	.0323 (.0593)	-.2867 (.1950)	.2309* (.1176)	.0311 (.1680)	.4032*** (.1202)
Difference between AI conditions (PB – PW)	.2037*** (.0599)	.1973*** (.0587)	.2049 (.2300)	.4403*** (.1072)	.7596*** (.0641)	-.0933 (.1527)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	72	36	36	72	36	36
Squared correlation	.8890	.4320	.2639	.8189	.9759	.7666
Pseudo R ²	.9810	.9862	.9618	.9872	.9936	.9907
BIC	395	226	170	6,796	1,932	2,150

Table A43: Regression winsorization results, Study 3

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.
The unit of observation is an ad-platform-age-gender combination.

P Visual Content Similarity

To quantify visual diversity across assets, we implemented ViViT, a video vision transformer model, to analyze and compare videos based on their lower-dimensional embeddings (Arnab et al., 2021). ViViT captures features for frame extraction, optical flow alignment, embedding generation, and similarity computation. To evaluate visual diversity, we used the 33 human video ads and randomly sampled 33 AI video ads (one per campaign) from Studies 1-3. We embedded all 66 videos into a 768-dimensional, real-valued vector using the ViViT model and calculated pairwise cosine similarity measures across video ads within each condition.

The pairwise cosine similarity results, visualized in Figure A7, reveal that AI video ads demonstrate significantly lower average within-group similarity ($M = .4760$) compared to human video ads ($M = .5343$, $t(1009) = -9.08$, $p < .001$).

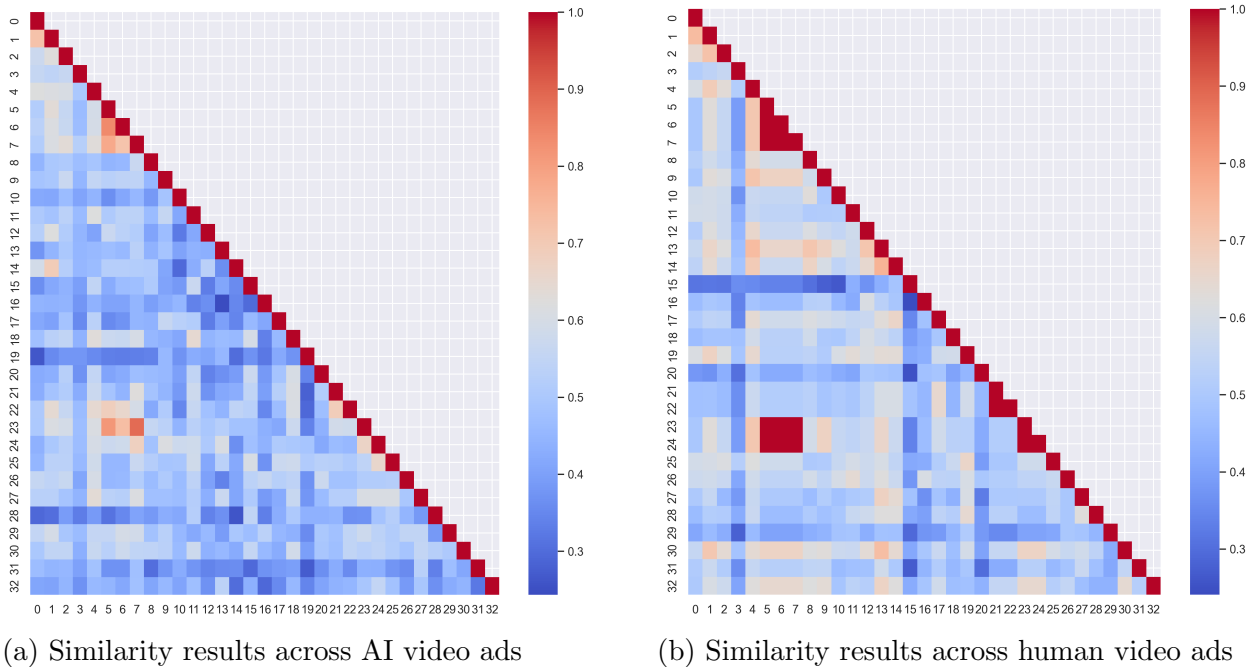


Figure A7: Similarity heatmaps for AI and human video ads

Notes: Typically, albums feature multiple tracks, resulting in video ads with identical packshots and consequently reduced visual diversity. While human-created assets demonstrate complete visual uniformity in these instances, MAGE maintains visual variation by implementing dynamically generated backgrounds and diverse parameter sampling across all assets.

To investigate if multiple AI videos within the same campaign influence performance

by speaking a similar visual language, we use the same ViViT-based embedding model and compute the average pairwise similarity between every video ad to all other video ads in the campaign. Tables A44, A45, and A46 present the results.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall CTR	TikTok CTR	Meta CTR	Overall ER	TikTok ER	Meta ER
<i>Video ads</i>						
Human (baseline)						
AI	.1509*** (.0418)	.2341*** (.0470)	-.1104 (.0818)	-.0558* (.0213)	.0159 (.0220)	-.1158*** (.0285)
Visual Similarity _z	.0213 (.0205)	.0226 (.0238)	.0223 (.0362)	.0102 (.0098)	.0021 (.0108)	.0183 (.0132)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	4,114	2,239	1,875	4,114	2,239	1,875
Squared correlation	.0233	.0120	.0140	.1756	.0943	.5925
Pseudo R ²	.9263	.9412	.8834	.985	.9876	.9860
BIC	9,771	5,865	3,939	47,572	19,071	22,873

Table A44: Regression results with visual similarity, Study 1

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall CTR	TikTok CTR	Meta CTR	Overall ER	TikTok ER	Meta ER
<i>Video ads</i>						
Human (baseline)						
AI: randomly sampled (RS)	.2438* (.0955)	.2557* (.1051)	.3875† (.2188)	.0259 (.0486)	.0970* (.0484)	-.0491 (.0788)
AI: expert selected (ES)	.1391 (.1064)	.1922 (.1219)	.1277 (.2189)	-.0072 (.0507)	.0546 (.0571)	-.0853 (.0742)
Visual Similarity _z	-.0224 (.0413)	-.0168 (.0462)	-.0542 (.0809)	.0417* (.0197)	.0452* (.0227)	.0426 (.0275)
Difference between AI conditions (RS – ES)	.1047† (.0622)	.0635 (.0680)	.2598* (.1208)	.0331 (.0272)	.0424 (.0286)	.0362 (.0392)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	1,992	1,048	944	1,992	1,048	944
Squared correlation	.00957	.00540	.0129	.1383	.0745	.2083
Pseudo R ²	.9033	.9186	.8478	.9804	.9819	.9797
BIC	3,207	2,146	1,088	17,958	7,698	9,683

Table A45: Regression results with visual similarity, Study 2

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
Dep. Var.:	Overall	TikTok	Meta	Overall	TikTok	Meta
	CTR	CTR	CTR	ER	ER	ER
<i>Video ads</i>						
Human (baseline)						
AI: predicted best CTR (PB)	.2514 (.1972)	.3558† (.1866)	.0263 (.3578)	.0200 (.2422)	.4394 (.3263)	.2001 (.2069)
AI: predicted worst CTR (PW)	.0306 (.1521)	.1249 (.1449)	-.1401 (.4108)	-.3688* (.1850)	-.2209 (.2553)	.2541 (.1971)
Visual Similarity _z	-.0389 (.0776)	-.0493 (.0752)	-.0495 (.1414)	.2567** (.0832)	.1297 (.0986)	.0558 (.0898)
Difference between AI conditions (PB – PW)	.2208** (.0769)	.2309** (.0751)	.1664 (.2311)	.3888*** (.0949)	.6603*** (.1060)	-.0541 (.1439)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	72	36	36	72	36	36
Squared correlation	.8887	.4398	.27843	.8778	.9759	.7703
Pseudo R ²	.9811	.9862	.9619	.9896	.9938	.9908
BIC	398.4	228.9	173.5	5,523	1,863	2,131

Table A46: Regression results with visual similarity, Study 3

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination.

Q Quantifying the Business Impact of MAGE

To quantify the potential economic impact of MAGE at scale, we performed a back-of-the-envelope production cost analysis. The entire music catalog of a large music label might include one million tracks. Assuming a music label wants to promote 3% of those music tracks per year with four video ads per campaign, then this would result in $N_{\text{assets}} \approx N_{\text{tracks}} \times P_{\text{promotion}} \times N_{\text{assets/campaign}} \approx 1,000,000 \times .03 \times 4 = 120,000$ video ads to be created. If these assets were manually created, the total cost would be $C_{\text{human}} = 120,000 \times \text{USD } 100 = \text{USD } 12,000,000$, which is prohibitively high, such that the labels probably do not make this investment. Using MAGE though, the asset production cost would decrease to $C_{\text{MAGE}} = 120,000 \times \text{USD } 1.75 = \text{USD } 210,000$, a reasonable fee making automated content creation with MAGE a highly cost-effective alternative, which in turn enables new marketing opportunities previously constrained by budget limitations. Furthermore, traditional productions of 120,000 video ads would require a large team of full-time designers; MAGE needs only one operator.

To quantify the business impact of segmentation beyond content creation cost savings, we simulate campaign performance across demographic segments. Across 1,680 simulations, we find a CTR increase of 21.5% on TikTok and 9.8% on Meta, see Table A47.

Platform:	TikTok				Meta			
Dep. Var.:	CTR (in %)			Uplift	CTR (in %)			Uplift
Model:	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Selection:	Predicted worst CTR	Randomly sampled	Predicted best CTR	(3) vs. (2)	Predicted worst CTR	Randomly sampled	Predicted best CTR	(7) vs. (6)
Female	.1379	.1649	.1993	20.8%	.0581	.0663	.0730	10.2%
18-24	.1772	.2037	.2360	15.8%	.0592	.0675	.0742	9.9%
25-34	.1299	.1569	.1912	21.8%	.0550	.0617	.0670	8.5%
35-44	.1228	.1485	.1826	23.0%	.0561	.0642	.0707	10.1%
45-54	.1253	.1538	.1861	21.0%	.0591	.0685	.0764	11.6%
55+	.1343	.1618	.2007	24.1%	.0611	.0696	.0770	10.6%
Male	.1429	.1708	.2087	22.2%	.0584	.0668	.0731	9.4%
18-24	.1819	.2096	.2442	16.5%	.0594	.0681	.0743	9.1%
25-34	.1349	.1628	.2012	23.6%	.0559	.0628	.0676	7.7%
35-44	.1277	.1540	.1921	24.8%	.0559	.0645	.0707	9.6%
45-54	.1308	.1599	.1963	22.8%	.0593	.0687	.0762	10.8%
55+	.1394	.1679	.2098	25.0%	.0613	.0701	.0767	9.5%
Overall	.1404	.1679	.2040	21.5%	.0582	.0666	.0731	9.8%

Table A47: Simulation results for AI-driven, segment-specific content creation

Notes: Based on N=1,680 simulations for the seven campaigns from Study 2. CTR prediction based on platform-specific models trained on performance data from Study 1. Predicted best (worst) CTR computed by Bayesian optimization to maximize (minimize) predicted CTR.

R Landing Page Engagement

When users click on ad links, they are directed to a Linkfire landing page,⁷ where they can either listen to the song directly on the page or navigate to connected streaming platforms such as Spotify or Apple Music. The landing page engagement rate (engagement divided by clicks, LP-ER) encompasses both on-page streaming activity and downstream clicks to streaming platforms. Table A48 presents the regression results for Studies 1-3. Demographic breakdowns by age or gender are not available for landing page visits. If such a pattern holds, the treatment effect might diminish further along the conversion funnel.

Study:	Study 1		Study 2		Study 3	
Platform:	TikTok	Meta	TikTok	Meta	TikTok	Meta
Dep. Var.:	LP-ER	LP-ER	LP-ER	LP-ER	LP-ER	LP-ER
<i>Video ads</i>						
Human (baseline)						
AI	-.5461 (.4586)	-.2914† (.1589)				
AI: randomly sampled			-1.726 (1.086)	.3290 (.3389)		
AI: expert selected			-1.002 (.7025)	.2090 (.3606)		
AI: predicted best CTR					-.6119 (1.127)	-1.343 (1.242)
AI: predicted worst CTR					-.5176 (1.082)	-.0394 (.9224)
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	No ¹	No ¹
<i>Fit statistics</i>						
Observations	99	125	27	39	36	31
Squared correlation	.1777	.485263	.2099	.1575	.0208	.0714
Pseudo R ²	.70566	.8009	.6737	.7696	.5692	.5823
Bayesian information criterion	299	597	38	107	69	77

Table A48: Regression results for landing page engagement, Studies 1-3.

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$.

¹ In Study 3, there is only one music track. Hence, there is no music track fixed effect.

⁷For an example landing page, see Linkfire, 2025.

S Genre

This section reports exploratory heterogeneity by music genre. Table A49 reports the genre-specific effects of AI video ads across the genres represented in the field studies.

Model:	(1)	(2)	(3)	(4)	(5)	(6)
	Overall	TikTok	Meta	Overall	TikTok	Meta
Dep. Var.:	(Pooled)	(Pooled)	(Pooled)	(Pooled)	(Pooled)	(Pooled)
	CTR	CTR	CTR	ER	ER	ER
<i>Video ads</i>						
Human (baseline)						
AI	.2601*** (.0700)	.3999*** (.0798)	-.1802† (.1091)	-.0341 (.0303)	.0312 (.0264)	-.0859* (.0438)
Electronic (baseline)						
AI × Classical	-.2990* (.1398)	-.4405** (.1605)	.1570 (.3472)	.0911 (.0632)	.0880 (.0812)	.0898 (.0820)
AI × Folk	-.2192 (.2141)	-.3277 (.2165)	.1420 (.3711)	-.1880† (.1067)	-.2533*** (.0545)	-.1307† (.0711)
AI × Jazz	-.2090 (.1744)	-.3198† (.1924)	.1514 (.3107)	-.0214 (.0636)	.0242 (.0824)	-.0476 (.0777)
AI × Pop	-.1178 (.0880)	-.1833† (.0993)	.1103 (.1761)	.0990* (.0471)	.1057** (.0404)	.0030 (.0767)
AI × Rock	-.1071 (.0803)	-.2072* (.0900)	.1624 (.1360)	.1250† (.0654)	.0730 (.0688)	.1076† (.0605)
<i>Controls</i>						
Audience age	Yes	Yes	Yes	Yes	Yes	Yes
Audience gender	Yes	Yes	Yes	Yes	Yes	Yes
<i>Fixed effects</i>						
Music track	Yes	Yes	Yes	Yes	Yes	Yes
Platform	Yes	No	No	Yes	No	No
<i>Fit statistics</i>						
Observations	6,178	3,323	2,855	6,178	3,323	2,855
Squared correlation	.0121	.0065	.0110	.1560	.0943	.3575
Pseudo R ²	.9278	.9420	.8852	.9818	.9849	.9841
BIC	13,419	8,242	5,211	83,930	34,154	37,310

Table A49: Regression results including genre, Studies 1-3

Notes: Robust standard errors are in parentheses. *** $p < .001$, ** $p < .01$, * $p < .05$, † $p < .1$. The unit of observation is an ad-platform-age-gender combination. Based on data from Studies 1-3.

Appendix References

- Angelopoulos, P., Lee, K., & Misra, S. (2024). Value Aligned Large Language Models. *SSRN Electronic Journal*. <https://doi.org/10.2139/ssrn.4781850>
- Arnab, A., Dehghani, M., Heigold, G., Sun, C., Lučić, M., & Schmid, C. (2021). ViViT: A Video Vision Transformer. <https://doi.org/10.48550/arXiv.2103.15691>
- Blattmann, A., Dockhorn, T., Kulal, S., Mendelevitch, D., Kilian, M., Lorenz, D., Levi, Y., English, Z., Voleti, V., Letts, A., Jampani, V., & Rombach, R. (2023). Stable Video Diffusion: Scaling Latent Video Diffusion Models to Large Datasets. <https://doi.org/10.48550/arXiv.2311.15127>
- Boegershausen, J., Cornil, Y., Yi, S., & Hardisty, D. J. (2025). On the persistent mischaracterization of Google and Facebook A/B tests: How to conduct and report online platform studies. *International Journal of Research in Marketing*. <https://doi.org/10.1016/j.ijresmar.2024.12.004>
- Braun, M., de Langhe, B., Puntoni, S., & Schwartz, E. M. (2024). Leveraging Digital Advertising Platforms for Consumer Research. *Journal of Consumer Research*, 51(1), 119–128. <https://doi.org/10.1093/jcr/ucad058>
- Braun, M., & Schwartz, E. M. (2024). Where A-B Testing Goes Wrong: How Divergent Delivery Affects What Online Experiments Cannot (and Can) Tell You about How Customers Respond to Advertising. *Journal of Marketing*. <https://doi.org/10.1177/00222429241275886>
- Burtch, G., Moakler, R., Gordon, B. R., Zhang, P., & Hill, S. (2025). Characterizing and Minimizing Divergent Delivery in Meta Advertising Experiments. <https://arxiv.org/pdf/2508.21251>
- Celebi, M. E. (2011). Improving the performance of k-means for color quantization. *Image and Vision Computing*, 29(4), 260–271. <https://doi.org/10.1016/j.imavis.2010.10.002>
- Chen, J., Chi, L., Xu, S., Ran, S., Peng, B., & Yuan, Z. (2025). HLLM-Creator: Hierarchical LLM-based Personalized Creative Generation. <https://arxiv.org/pdf/2508.18118>
- Chen, Z., & Chan, J. (2024). Large Language Model in Creative Work: The Role of Collaboration Modality and User Expertise. *Management Science*, 70(12), 9101–9117. <https://doi.org/10.1287/mnsc.2023.03014>
- Cian, L., Krishna, A., & Elder, R. S. (2014). This Logo Moves Me: Dynamic Imagery from Static Images. *Journal of Marketing Research*, 51(2), 184–197. <https://doi.org/10.1509/jmr.13.0023>
- Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2. ed.). Psychology Press. <http://www.loc.gov/catdir/enhancements/fy0731/88012110-d.html>
- Daviet, R., & Nishimura, Y. (2026). Leveraging Generative Artificial Intelligence to Create Visual Content in Digital Advertising. *Marketing Science*. <https://doi.org/10.1287/mksc.2024.1130>
- Dubé, J.-P., & Xu, A. (2026). Large Language Models and Creative Content Design: a case study of email marketing at Wine Access. *Quantitative Marketing and Economics*, 24(1). <https://doi.org/10.1007/s11129-025-09303-9>
- Efthymiou, F., Hildebrand, C., de Bellis, E., & Hampton, W. H. (2024). The Power of AI-Generated Voices: How Digital Vocal Tract Length Shapes Product Congruency and

- Ad Performance. *Journal of Interactive Marketing*, 59(2), 117–134. <https://doi.org/10.1177/10949968231194905>
- Esser, P., Kulal, S., Blattmann, A., Entezari, R., Müller, J., Saini, H., Levi, Y., Lorenz, D., Sauer, A., Boesel, F., Podell, D., Dockhorn, T., English, Z., Lacey, K., Goodwin, A., Marek, Y., & Rombach, R. (2024). Scaling Rectified Flow Transformers for High-Resolution Image Synthesis. <https://doi.org/10.48550/arXiv.2403.03206>
- Exner, Y., Hartmann, J., Netzer, O., & Zhang, S. (2025). AI in disguise - How AI-generated ads' visual cues shape consumer perception and performance. *SSRN Electronic Journal*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5096969
- Fang, L., Yuan, Z., Zhang, K., Donati, D., & Sarvary, M. (2025). Generative AI and Firm Productivity: Field Experiments in Online Retail. <https://doi.org/10.2139/ssrn.5584850>
- Gordon, B. R., Moakler, R., & Zettelmeyer, F. (2023). Close Enough? A Large-Scale Exploration of Non-Experimental Approaches to Advertising Measurement. *Marketing Science*, 42(4), 768–793. <https://doi.org/10.1287/mksc.2022.1413>
- Hartmann, J., Exner, Y., & Domdey, S. (2025). The power of generative marketing: Can generative AI create superhuman visual marketing content? *International Journal of Research in Marketing*, 42(1), 13–31. <https://doi.org/10.1016/j.ijresmar.2024.09.002>
- Heitmann, M., Jansen, T. P., Reisenbichler, M., & Schweidel, D. A. (2025). Picture Perfect: Engaging Customers with Visual Generative AI. *Journal of Marketing*. <https://doi.org/10.1177/00222429251356993>
- Henderson, C. M., Mazodier, M., & Sundar, A. (2019). The Color of Support: The Effect of Sponsor–Team Visual Congruence on Sponsorship Performance. *Journal of Marketing*, 83(3), 50–71. <https://doi.org/10.1177/0022242919831672>
- Hou, J., Wang, L., Wang, G., Wang, H. J., & Yang, S. (2025). The Double-Edged Roles of Generative AI in the Creative Process: Experiments on Design Work. *Information Systems Research*. <https://doi.org/10.1287/isre.2024.0937>
- Hunt, J., & Kerr, D. (2025). The hybrid formula: The essential role of human creativity in AI-generated advertising. *Journal of Brand Strategy*, 14(1), 30. <https://doi.org/10.69554/NBYX9041>
- Jakesch, M., Hancock, J. T., & Naaman, M. (2023). Human heuristics for AI-generated language are flawed. *Proceedings of the National Academy of Sciences of the United States of America*, 120(11), e2208839120. <https://doi.org/10.1073/pnas.2208839120>
- Kapoor, A., & Kumar, M. (2025). Frontiers: Generative AI and Personalized Video Advertisements. *Marketing Science*, 44(4), 733–747. <https://doi.org/10.1287/mksc.2023.0494>
- Lee, H., Todri, V., Adamopoulos, P., & Ghose, A. (2025). The Impact of Visual Generative AI on Advertising Effectiveness. <https://doi.org/10.2139/ssrn.5638311>
- Linkfire. (2025). <https://www.linkfire.com/>.
- OpenAI, Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F. L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., Avila, R., Babuschkin, I., Balaji, S., Balcom, V., Baltescu, P., Bao, H., Bavarian, M., Belgum, J., ... Zoph, B. (2023). GPT-4 Technical Report. <https://doi.org/10.48550/arXiv.2303.08774>

- Reisenbichler, M., Reutterer, T., Schweidel, D. A., & Dan, D. (2022). Frontiers: Supporting Content Marketing with Natural Language Generation. *Marketing Science*, 41(3), 441–452. <https://doi.org/10.1287/mksc.2022.1354>
- Sella, F., Raz, G., & Cohen Kadosh, R. (2021). When randomisation is not good enough: Matching groups in intervention studies. *Psychonomic Bulletin & Review*, 28(6), 2085–2093. <https://doi.org/10.3758/s13423-021-01970-5>
- Smith, R. E., MacKenzie, S. B., Yang, X., Buchholz, L. M., & Darley, W. K. (2007). Modeling the Determinants and Effects of Creativity in Advertising. *Marketing Science*, 26(6), 819–833. <https://doi.org/10.1287/mksc.1070.0272>
- To, R. N., Wu, Y.-C., Kianian, P., & Zhang, Z. (2025). When AI Doesn't Sell Prada: Why Using AI-Generated Advertisements Backfires for Luxury Brands. *Journal of Advertising Research*, 65(2), 202–236. <https://doi.org/10.1080/00218499.2025.2454120>
- Yoo, C. Y. (2007). Implicit Memory Measures for Web Advertising Effectiveness. *Journalism & Mass Communication Quarterly*, 84(1), 7–23. <https://doi.org/10.1177/107769900708400102>
- Zhang, S., Lee, D., Singh, P. V., & Srinivasan, K. (2022). What Makes a Good Image? Airbnb Demand Analytics Leveraging Interpretable Image Features. *Management Science*, 68(8), 5644–5666. <https://doi.org/10.1287/mnsc.2021.4175>