GDG PROJECT - TEAM A
Behavior & Content Simulation for Social Media Marketing

***Abstract*—In the contemporary digital marketing landscape, success is fundamentally contingent upon an organization's capacity to generate and deploy content that achieves instantaneous and profound resonance with its target audience. This project directly addresses this strategic challenge by leveraging state-of-the-art Artificial Intelligence (AI) to create a closed-loop content intelligence ecosystem. Every granular component of a social media asset—including the linguistic style of the caption, the visual composition of the image, the temporal positioning of the post, and embedded metadata—constitutes a critical data signal that drives user perception, engagement, and long-term brand equity. The overarching aim is to establish sophisticated AI systems capable of two core functions: simulating granular audience behavior (prediction) and generating highly optimized, high-performing marketing content (generation) within a cohesive, unified pipeline.**

***Index Terms*—Artificial Intelligence, Social Media, Content Marketing, Predictive Modeling, Generative AI, Multimodal Learning, Machine Learning**

I. Introduction: The Need for Predictive Content Intelligence

This project directly addresses the strategic challenge of achieving profound resonance with a target audience by leveraging state-of-the-art Artificial Intelligence (AI) to create a closed-loop content intelligence ecosystem. Every granular component of a social media asset—including the linguistic style of the caption, the visual composition of the image, the temporal positioning of the post, and embedded metadata—constitutes a critical data signal that drives user perception, engagement, and long-term brand equity. The overarching aim is to establish sophisticated AI systems capable of two core functions: simulating granular audience behavior (prediction) and generating highly optimized, high-performing marketing content (generation) within a cohesive, unified pipeline. This capability is paramount for clients utilizing the Adobe Experience Cloud, facilitating drastic improvements in campaign Return on Investment (ROI) via data-driven personalization.

Our work is structured around two tightly integrated, foundational tasks:

1) **Behavior Simulation (Task 1)**: Development of a robust regression model to predict anticipated social media engagement (specifically, the number of likes) based on the comprehensive multimodal context of a post.
2) **Content Simulation (Task 2)**: Creation of a generative system for producing highly optimized tweet text that is both contextually relevant to the visual media and metadata, and conditioned on a desired engagement profile.

The successful integration of these two tasks validates a self-improving feedback mechanism: first, discerning the latent factors that drive engagement, and second, optimizing content creation based on that predictive knowledge.

II. Problem Statement and Technical Objectives

The official Adobe challenge defined two specific and technically challenging subtasks designed to evaluate both predictive modeling and generative AI capabilities on proprietary data.

A. Task 1 - Behavior Simulation (Predictive Regression)

- **Goal**: To accurately estimate the numerical volume of likes a tweet will accrue, given its full contextual feature set. This is a challenge in heteroscedastic regression due to the power-law distribution of the target variable.
- **Input Modalities**: Structured metadata (company, username, timestamp), Linguistic content (tweet text), and Visual context (media URLs).
- **Output**: A continuous, non-negative numeric prediction of expected likes, evaluated primarily using Root Mean Squared Error (RMSE).

B. Task 2 - Content Simulation (Generative Multimodal Translation)

- **Goal**: To synthesize high-quality, brand-consistent, and visually-grounded tweet text. The model is essentially performing a constrained, conditional text generation, where the condition is the multimodal context and a target engagement metric.
- **Input Features**: Full multimodal context (Company, Username, Media URLs), Temporal factors (Timestamp), and an Affective Feature (a target number of likes).
- **Output**: Generated natural language tweet text, evaluated using standard Natural Language Generation (NLG) coherence and quality metrics.

III. Task 1 - Behavior Simulation: Predicting Engagement

A. Objective and Rationale

The primary objective was to develop a resilient and accurate regression model capable of forecasting a social media post's engagement level. This required integrating and

weighing signals from three disparate modalities: textual content, structured brand/temporal metadata, and complex visual media. The insights gained are crucial for understanding the latent factors—such as visual composition, brand tone, or optimal posting time—that drive user reactions.

B. Dataset Overview and Challenges

The dataset comprised approximately **300K tweets** from enterprise-level brand Twitter accounts spanning 2018-2023. Features included date, time, company, username, media URL, tweet text, and the target variable, **likes** (continuous numeric).

The data presented significant inherent challenges typical of real-world social media data:

TABLE I
KEY OBSERVATIONS ON TARGET VARIABLE DISTRIBUTION

Metric	Value / Observation
Mean Likes	718.39
Std. Deviation	3866.48
Median	73
Skewness	28.48
Kurtosis	1424.22

Key Observations from Exploratory Data Analysis (EDA):

1) **Power-Law Distribution**: The engagement distribution resembled a classic power-law curve, where the vast majority of posts received less than 100 likes, while a small, select group (the "long tail") exceeded 10,000 likes.
2) **Multimodal Influence**: Posts containing media consistently showed a statistically significant higher median number of likes compared to text-only posts, confirming the indispensable role of visual content.
3) **Brand and Temporal Bias**: Engagement was strongly influenced by specific brand identities and temporal factors, peaking during weekends or major promotional months.

C. Data Pre-processing and Feature Engineering

The preprocessing pipeline focused on standardizing inputs and deriving rich features.

1) **Data Cleaning**:
   - Broken and non-existent media URLs were dropped to ensure feature integrity.
   - Timestamp formats were standardized to simplify temporal feature extraction.
2) **Engagement Bucketing**: The heavily skewed target variable (*likes*) was analyzed by forming discrete engagement buckets (e.g., 0-10, 10-100, 100-1K, etc.) to inform potential future hybrid modeling strategies.
3) **Feature Extraction and Embedding**:
   - **Metadata Features**: Temporal features like hour, weekday, and month were derived. Categorical fields (*company*, *username*) were label-encoded.
   - **Text Embedding**: The tweet text was processed using **DistilBERT** to generate high-dimensional, semantically rich vector embeddings. DistilBERT was chosen over classic TF-IDF/unsupervised BERT for its balance of performance and computational efficiency.
   - **Media Embedding**: For visual grounding, the **BLIP-2** model was used to generate image embeddings. These embeddings were designed to fuse visual and linguistic understanding, providing a semantic representation of the media content.

These robust preprocessing steps established a clean, multimodal foundation for all subsequent modeling efforts.

D. Approach Evolution and Evaluation

Our strategy involved an iterative progression, starting with a text-only baseline and progressively integrating the computationally expensive multimodal features.

1) Approach 1 - Text + Metadata Regression Baseline (Without Media): **Objective**: To establish a lower bound baseline using only textual and structured features. This approach deliberately ignored the media URL feature to quantify the performance cost of excluding visual context. A comprehensive suite of traditional machine learning and boosting models was tested (Ridge, Lasso, ElasticNet, Random Forest, XGBoost, LightGBM, CatBoost, Extra Trees, Gradient Boosting).

TABLE II
APPROACH 1 - TEXT/METADATA BASELINE PERFORMANCE (RAW LIKES)

Model	MAE (Raw Likes)	R^2 (Raw Likes)
Ridge Regression (Best)	525.0	0.1698
XGBoost	543.4	0.1047
CatBoost	560.6	0.0841

Conclusion: The low R^2 value confirmed that linguistic style and metadata alone explain very little of the engagement variance. This highlighted the critical, indispensable need for a strong visual context signal.

2) Approach 2 - Fused Multimodal Regression (Final Evaluated Model): **Objective**: To achieve superior performance by fully integrating the semantic embeddings from text and media.

Implementation and Training Scope:

- **Embedding Fusion**: The DistilBERT text embeddings, BLIP-2 media embeddings, and encoded metadata vectors were concatenated into a single, comprehensive feature vector.
- **Training Subset**: Due to computational constraints and the high rate of broken video/GIF URLs, this approach was trained and evaluated on a focused **3K sample subset** consisting only of image-based posts (videos and GIFs were explicitly removed to ensure smooth and complete BLIP-2 inference).
- **Target Transformation**: The target variable (*likes*) was **log-transformed** before training to mitigate the effects of

the extreme right skewness and stabilize gradient-based models (XGBoost, LightGBM). The evaluation metrics were computed on this transformed scale.

TABLE III
APPROACH 2 - MULTIMODAL XGBOOST PERFORMANCE (LOG-TRANSFORMED)

Model	RMSE (Log)	MAE (Log)	$\mathbf{R^2}$ (Raw)
XGBoost (Multimodal)	1.023	0.82	-0.046

Analysis and Interpretation:

- **Scaling Context**: An MAE of 0.82 on a scale where the target range is tightly normalized is an exceptionally high error value, indicating the model's prediction error is substantial relative to the target's standard deviation.
- **Raw $\mathbf{R^2}$ Failure**: The R^2 value ($\mathbf{-0.046}$) is reported on the original, raw target scale for direct comparison to the mean baseline. A negative R^2 is the most critical finding: it definitively shows that the complex multimodal model, despite using sophisticated features, performs slightly worse than a simple baseline that predicts the average number of likes.

Conclusion: This failure is primarily attributed to the insufficient and highly curated 3K image-only subset. The model, trained on such a small, isolated sample, lacked the generalizability to perform effectively on the broader test distribution. The high error on the normalized scale, coupled with the negative R^2, confirms the numerical instability and poor predictive power in this specific implementation.

IV. TASK 2 - CONTENT SIMULATION: GENERATIVE COPYWRITING

A. Objective and Context

The objective of Task 2 was to move beyond prediction and into generation: creating a system that mimics a high-performing social media copywriter. The model had to generate tweet text that was not only grammatically correct but also visually grounded, stylistically aligned with the brand, and optimized for engagement, using the context derived from the media and metadata.

- **Input Features**: Company, Username, Timestamp, Target Likes, and Media URLs/Captions.
- **Output**: Generated Tweet Text.
- **Embeddings Used**: BLIP-2 for media (via captioning) and DistilBERT for context features.

B. Approach Evolution

1) Approach 1 - LLM Fine-tuning Without Media: **Method**: Fine-tuned the **Mistral-7B-Instruct** model on the metadata (brand, time, target likes) but deliberately without any media context or captions.

Observation: The model produced extremely poor, repetitive, or generic outputs, such as "Check out our new product today!" or "Visit our site for more!".

Reasoning: Tweets are fundamentally visually grounded in the social media ecosystem. Without image/video semantics, the language generation lacked relevance and failed to generalize across diverse brands or topics, resulting in BLEU/ROUGE scores near zero.

2) Approach 2 - Florence-2 Captioning + Mistral-7B-Instruct v0.2 Fine-Tuning (Final Pipeline): **Motivation**: To achieve high-quality, semantically rich multimodal grounding for both images and videos before generation.

Architecture:

1) **Media Understanding**: The **Florence-2 Vision-Language Model** was used to generate high-quality, detailed captions.
2) **Video Handling**: For video posts, a robust method of **entropy-based key-frame selection** was implemented to extract up to five frames with maximal information content. The captions from these frames were then fused into a single, comprehensive scene description.
3) **Dataset Formation**: Each training sample was structured as: {metadata + Florence-2 captions} $\rightarrow$ {reference tweet text}.
4) **Model Fine-Tuning**: The **Mistral-7B-Instruct-v0.2** base model was fine-tuned using **LoRA** (Low-Rank Adaptation) for parameter-efficient training, minimizing cross-entropy loss.
5) **Post-Processing**: A post-processing step was added to clean the final generated tweet text, specifically removing residual artifacts like hashtags or extra tokens that the LLM sometimes injected.

Why this worked better: Florence-2 provided semantically rich, low-noise captions, which, when combined with metadata, taught the Mistral model a brand-consistent tone and posting style, leading to diverse and visually descriptive outputs.

C. Final Pipeline Summary

TABLE IV
TASK 2 - FINAL CONTENT SIMULATION PIPELINE

Stage	Component
1	Florence-2 Captioner
2	Entropy-based Frame Selection
3	Caption Fusion Layer
4	Mistral-7B-Instruct v0.2 (LORA Fine-Tuned)
5	Post-Processing

V. RESULTS, EVALUATION & CONCLUSION

A. Task 1: Behavior Simulation (Prediction)

The final evaluation compares the performance of the non-multimodal baseline (Approach 1) versus the log-transformed, fused multimodal model (Approach 2) trained on the 3K image-only subset.

TABLE V
TASK 1 PERFORMANCE COMPARISON

Metric	Approach 1: Ridge Regression (Text + Metadata)	Approach 2: Fused Multimodal XGBoost (Image-Only, Log-Transformed)
Training Scale	Full 300K Dataset (Raw Likes)	Small 3K Image Subset (Log Likes)
Media Included	No	Yes (BLIP-2 Embeddings)
RMSE	$\sim$550.0 (Est. based on MAE/R^2)	1.023
MAE	525.0	0.82
R^2	0.1698	-0.046

1) Interpretation of Task 1 Results:

1) **Metric Scale Disparity**: The dramatic difference in MAE (525.0 vs. 0.82) confirms the necessity of log-transforming the highly skewed target variable (*likes*). Approach 1, trained on the raw scale, inherently prioritized reducing the error on the rare, massive outliers, leading to high overall MAE. Approach 2, trained on the log scale, aimed to minimize the relative error, which is more robust for general prediction.
2) **Multimodal Impact**: The **negative** R^2 in Approach 2 is highly indicative of severe overfitting or mis-calibration on the small 3K training subset. While the inclusion of BLIP-2 features is theoretically superior, the restricted sample size was insufficient to allow the complex multimodal model (XGBoost) to learn generalizable patterns, causing it to fail on the test set.
3) **Conclusion**: The results forcefully underscore two points: (a) Log-transformation is necessary for skewed engagement data, and (b) Multimodal features require a full-scale, robust dataset to prevent overfitting; running complex VLM pipelines on a tiny, curated subset is insufficient for reliable performance.

B. Task 2: Content Simulation (Generation)

The generative model, utilizing Florence-2 captions and the LoRA-fine-tuned Mistral-7B-Instruct, was evaluated against three standard NLG metrics.

1) Interpretation of Task 2 Results: The results demonstrate a substantial, measurable, and highly significant quantitative uplift across all relevant NLG metrics following the integration of the Florence-2 visual grounding layer.

1) **Perplexity Reduction**: The drop in Perplexity from 10.12 to 3.6349 indicates that the model is now dramatically more confident and fluent when generating text, as the multimodal context allows it to assign a higher probability to the correct sequence of words. This is the hallmark of a successful conditional generation model.
2) **ROUGE Score Validation**: The high ROUGE scores (especially the $\mathbf{0.57}$ for ROUGE-4) are exceptionally strong for a generative task. ROUGE-4 measures the overlap of four-word sequences (4-grams), indicating that the generated text not only contains the right keywords but also correctly reproduces the longer, complex phrases and stylistic patterns present in the reference marketing copy. This confirms the model's ability to transition from producing generic text (Negligible ROUGE scores in Approach 1) to generating contextually relevant, human-like marketing copy.

C. Overall Project Conclusion

The dual-task framework successfully validated the core concept of a closed-loop content simulator, establishing robust and instructive baselines for both prediction and generation.

1) **Visual Context is Indispensable**: The failure of pure text-based models in both Task 1 ($\mathbf{R^2 < 0.17}$) and Task 2 (Near-Zero BLEU/ROUGE) proves that social media engagement is an inherently **multimodal** and affective phenomenon. Integration of advanced Vision-Language Models (BLIP-2/Florence-2) is a non-negotiable requirement for achieving practical utility.
2) **Data Distribution Strategy**: The challenge of the power-law distribution in engagement data necessitates sophisticated target transformations ($\mathbf{log1p}$) and, ideally, a planned bucketed modeling approach to stabilize gradient-based predictors against extreme outliers.
3) **Efficiency and Generative Quality**: The combination of a high-fidelity VLM for grounding (Florence-2) and parameter-efficient fine-tuning ($\mathbf{LoRA}$) proved to be the optimal strategy for the constrained, conditional generation task, validating the pipeline's ability to produce high-quality, visually grounded content.

This project validates the multimodal pipeline for both forecasting engagement and creating optimized content, offering a strong baseline for future AI-assisted marketing tools.

VI. OBSERVATIONS & LEARNINGS

A. Multimodal Dependency and Feature Weighting

Pure text-based approaches fundamentally fail when dealing with social media engagement, which is an inherently visual and emotional reaction. The integration of high-quality media embeddings (BLIP-2) or captions (Florence-2) provided the most significant feature uplift. For instance, SHAP analysis on early models indicated that media presence and brand identity were exponentially more influential than text complexity or length.

TABLE VI
TASK 2 PERFORMANCE METRICS

Metric	Measure	Approach 1: Text-Only Mistral 7B (Baseline)	Final Approach: Florence-2 + Mistral v0.2
Perplexity	(Lower is Better)	10.12	**3.6349**
ROUGE-1 (Unigram Overlap)	(Higher is Better)	Negligible	**0.29**
ROUGE-2 (Bigram Overlap)	(Higher is Better)	Negligible	**0.41**
ROUGE-4 (4-gram Overlap)	(Higher is Better)	Negligible	**0.57**

B. The Challenge of Data Skewness

The power-law engagement distribution proved to be the single greatest hurdle for Task 1. Simple regressors could not reliably capture the non-linear, brand-specific trends. The proposed bucket-based strategy is the theoretically correct solution, as it aims to reduce target variance within localized ranges, thereby stabilizing gradient-based learning.

C. Model Fine-Tuning and Generative Quality

Instruction-tuned LLMs like Mistral-Instruct adapted faster and generalized better when guided by contextual input. The LoRA fine-tuning method was not only practical for computational constraints but also highly effective at steering the model's output toward a specific 'copywriting' style, rather than relying on less efficient full-parameter fine-tuning.

D. The Pipeline Cruft: Media Accessibility

The high incidence of broken or inaccessible media URLs across the full corpus substantially restricted the ability to train Approach 2 on the entire 300K dataset. This issue underlines the critical need for robust data caching, pre-fetching, and fallback mechanisms in industrial-scale multimodal pipelines.

E. Evaluation Limitations

While standard metrics were used, BLEU/ROUGE/CIDEr for Task 2 often fail to capture stylistic nuance, brand voice, or emotional impact. The ultimate measure of success for generated marketing copy is actual engagement (likes/clicks), suggesting that future work should incorporate engagement-based scoring (like RLHF) or subjective human evaluations.

VII. FUTURE SCOPE

A. Enhanced Media Understanding

- **Direct Feature Embedding**: Move beyond BLIP-2/Florence-2 captioning and integrate direct visual/video features from state-of-the-art vision models (e.g., Flamingo, GPT-4V) to capture richer, lower-level semantics without relying on the potentially lossy intermediate captioning step.
- **Full Video Encoding**: Implement specialized video processing to create temporal-aware embeddings that track scene changes, rather than relying only on a few key frames.

B. Full-Scale Dataset Training

The primary next step for Task 1 is to resolve the media accessibility issues and train the Fused Multimodal Regression model on the entire 300K dataset. This will address the acute under-representation in high-engagement buckets and solve the generalization failure observed in Approach 2.

C. Joint Task Optimization and Feedback Loop

Combine both prediction and generation into a truly iterative pipeline: Generate a new tweet $\rightarrow$ Predict its engagement $\rightarrow$ Reinforce the generation based on the predicted score. This forms a self-improving content simulator that learns to generate content that its own internal predictor forecasts as high-performing.

D. Explainability and Visualization

Implement model explainability techniques like **SHAP** (SHapley Additive exPlanations) or LIME to provide marketers with transparency. This would explicitly reveal which visual, textual, or temporal factors drove a high predicted engagement score, turning the black-box model into an actionable tool for creative guidance.

REFERENCES

[1] Project Report
[2] Information from source