SmartEdit: Editing-driven Engagement Prediction and Enhancement of Short-Videos

Saumya Gupta[1]*, Ishita Dasgupta[2], Stefano Petrangeli[2], Somdeb Sarkhel[2]

*work carried out during summer 2024 internship at Adobe. [1]Stony brook University [2]Adobe Research

Motivation

I. Creators lack tools to predict short video engagement before publication
II. Limited editing experience makes social media video optimization challenging
III. Need for automated video refinement to boost engagement potential

Challenges

Data & Annotation: Lack of fine-grained labels linking editing actions (e.g., story flow, music, text, transitions) to engagement

Evaluation & Generalizability: Existing engagement prediction methods depend on platform-specific, post-publication metrics whose insights are not generalizable

Feedback & Interpretability: Current SOTA can understand actions and objects (pixel-level) information in a short-video but cannot ground editing operations to user behavior (engagement) for efficient editing-driven interpretable feedback

SmartEdit

We propose SmartEdit, which analyses short videos on the basis of video editing operations.

It is able to

1. Predict an **engagement score**
2. Provide a **ranking of editing operations** *(Edit Signals)* that positively or negatively influence engagement
3. Suggest **interpretable feedback** for creators to enhance content to improve engagement

References

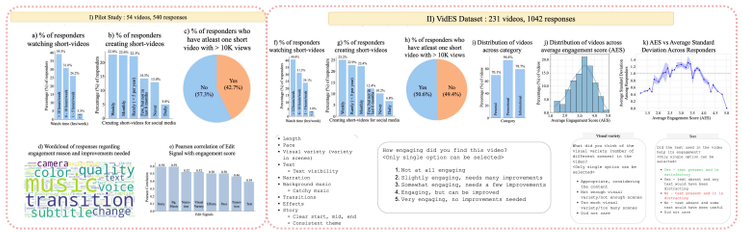
Dasong Li et al., "Delving deep into engagement prediction of short videos," in ECCV 2025.

Haoning Wu et al., "Q-align: Teaching lmms for visual scoring via discrete text-defined levels," in ICML 2024

Zhanyu Wang et al., "Gpt4video: A unified multimodal large language model for lnstruction-followed understanding and safety-aware generation," in ACM lCM 2024.

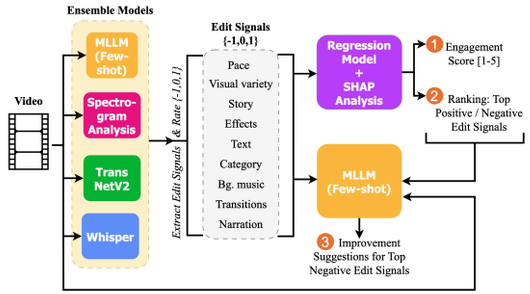
Proposed Approach

I. Curate a **novel dataset, VidES**, containing short-videos, engagement scores, and detailed human evaluations of Edit Signals, establishing a foundation for data-driven approaches to short-video refinement.



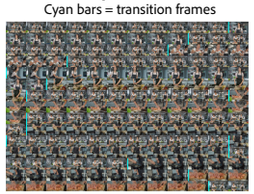
Validation study & full-scale dataset collection along with snippets of survey questions asked to the users

II. Propose **SmartEdit**, a **two-stage framework** that decomposes short-video engagement prediction into interpretable video editing operations (Edit Signals):

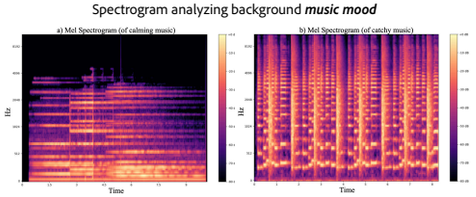


- ***Short-video Engagement Prediction***: Random forest regressor trained on VidES data to predict an engagement score based on extracted edit signals
- ***Interpretable feedback for improvement***: a) SHAP analysis to rank top engagement driving and harming Edit Signals, and b) few-shot MLLM to generate actionable improvement suggestions grounded in structured user feedback.

Visualizing ***transition detection.***
Cyan bars = transition frames



Spectrogram analyzing background ***music mood***



Results

1. SmartEdit ***outperforms baselines*** (random, Q-Align, GPT-4o zero/few-shot) in ***engagement prediction*** and ***feedback accuracy***
2. Ablation studies show the importance of both good and bad examples, and reasoning in prompting
3. Qualitative examples demonstrate iterative improvement of videos using SmartEdit's feedback

Performance on Engagement Prediction & Improvement Feedback

Method	ES Decomp.	Engagement Prediction MAE ↓	RMSE ↓	clfAcc ↑	Improvement Feedback Accuracy ↑	F1 ↑
Random model	✗	1.45	1.77	0.21	0.27	0.43
Q-Align [20]	✗	0.79	0.97	0.39	N/A	N/A
GPT-4o (Zero-shot)	✗	1.19	1.39	0.13	0.36	0.54
GPT-4o (Few-shot)	✗	1.02	1.23	0.28	0.41	0.57
GPT-4o (Few-shot)	✓	0.75	0.88	0.38	0.45	0.61
SmartEdit	✓	**0.64**	**0.84**	**0.49**	**0.51**	**0.67**

Ablation study testing different backbone for engagement prediction model

Regression Model	MAE ↓	RMSE ↓	clfAcc ↑
XGBoost	1.15	1.33	0.28
Multi-Layer Perceptron (MLP)	0.91	1.18	0.37
Random Forest Regressor	**0.64**	**0.84**	**0.49**

Ablation study on feedback improvement

#Good	#Bad	Reasoning?	Accuracy ↑	F1 ↑
0	4	✗	0.44	0.59
0	4	✓	0.47	0.63
2	2	✗	0.45	0.62
2	2	✓	**0.51**	**0.67**