A Four-Modality Fusion Method of Crowdfunding Prediction

Zijian ZHANG

Abstract

The goal of crowdfunding ex-ante prediction is to extract the most relevant information about the project in the context of crowdfunding to form reasoning. Unlike ex-post prediction, the ex-ante prediction does not rely on information obtained after the project is released. However, existing multimodality method fail to represent and align long sequences in the context of crowdfunding effectively, resulting in information noise and decreased prediction performance. To address this issue, we introduce a four-modality fusion method, which can realize modular extraction of multiple modalities to adapt to large-scale modality missing datasets and achieve semantic alignment and fusion in video frame space.

In addition, we propose two large-scale multimodality crowdfunding datasets, Kick30 and Kick60, each containing 30,000 and 68,000 crowdfunding projects respectively. The former also includes video frame features extracted by TimeSformer and textual sequence features extracted by BERT. The experiment on Kick30 demonstrate the superiority of our method, achieving state-of-the-art performance in 10/15 of the category. Extensive modality ablation study on both dataset shows the importance of visual and textual information in the reasoning process. Our code and dataset are published available at *https://github.com/zjzhang1999/cfmp.github.io*

1 Introduction

Crowdfunding research has witnessed a rapid advancement in the past few years, where multiple data with increasingly larger scale have been developed to continuously push the state-of-the-art on crowdfunding prediction tasks. According to the availability of data, one of the crowdfunding prediction tasks is based on prior information, such as project introduction and entrepreneur background before project release; the other is based on the post hoc, there are comments and messages from project supporters, supporters. There is no doubt that the former has greater reference value for platforms and entrepreneurs. Therefore, the crowdfunding prediction discussed in this paper refers to the ex-ante prediction, which will not be distinguished later.

The earliest research on crowdfunding prediction was based on single text modalities or images or meta data. However, with the emergence of new large-scale multimodal data sets and the improvement of GPU fast computing performance, the number of studies using two or more modes for crowdfunding prediction has begun to increase. Multimodal prediction normally achieves better performance due to the use of more information fusion. The goal of multimodal learning is to map the data of different modes such as speech, images, and text into a unified semantic space. There are always two important challenges in multimodal machine learning. That is, the representation and alignment of different modalities.

In the latest research, the i-Code (Yang et al.,2023) released by Microsoft can replace the single-modality encoder with any pre-trained model through the designed modular framework, which provides high-quality context representation for the fusion network for more effective multi-modal understanding. Meta's ImageBind (Girdhar et al.,2023), is trained with data from six modalities, including vision, text, audio, depth, thermal and IMU data. Such practice is equivalent to training five bi-modal models, each of which is a combination of the other five modality and the visual modality. In this way, all modalities are aligned to the representation space of image modality, so that all modalities can be unified into a shared representation space.

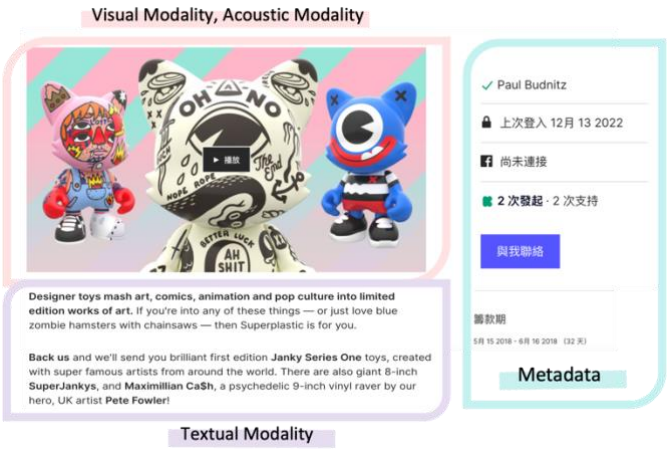


Figure I. An original data sample in Kick30.

Inspired by this, we propose a generic and compute-efficient fusion approach, which learn a joint embedding across four modalities- text, video, audio, and meta data. It does not need datasets where all modalities co-occur with each other. Instead, after obtaining available representation through modular pre-trained feature coding, we align each modality's embedding to video embeddings through cross-attention blocks and thus achieve modality fusion and semantic unity. In the decoder, the structured metadata will be added to the classifier to strengthen the understanding of crowdfunding context.

The approach is tested on Kickstarter30k, a crowdfunding dataset that contains 30,000 video-text pairs we independently collected. The experiments show that our model achieves better performance in predicting the success of 15 categories of crowdfunding projects than the benchmark.

The contribution of this research summarized as follows:

- This work is the first to apply four modalities in crowdfunding reasoning task.
- Our method achieves weak alignment of multiple crowdfunding modalities to facilitate reasoning at a lower cost. Our method achieves new state-of-the-art performance on the ex-ante crowdfunding benchmark.
- We collected Kick60k, a crowdfunding multimodal dataset with 60,000, 15-category video-text-meta projects. We publish it along with our method on GitHub for the later research.

2 Related Work

This section reviews recent works of crowdfunding (CF) prediction by unimodal and multimodal models, as well as summaries the main challenges in CF prediction.

2.1 CF Prediction and Challenge

CF Prediction with Unimodal. Single modality in crowdfunding has been widely used to examine the factors influencing the success of crowdfunding. Specifically, meta data and text are the most two widely used as its simplicity of sampling and interpretation. Recent Studies (Kaminski.,2020; Liang et al., 2019) have shown that crowdfunding success can be affected by three meta data involved in crowdfunding: creator, project, and platform. For creator, the credit status, entrepreneurial experience, social network, and other information of entrepreneurs are all important references for investors to make investment decisions. For project, the target, category, duration, as well as backers are often used as conditions for inference and explanation. For platform, the research showed that platform design, interface layout and human-computer interaction may also affect the results of crowdfunding.

Some prediction methods only use text modality achieves stronger performance than only use meta data and although they lack clear interpretability. Those studies (Lee et al.,2022; Kaminski.,2020; Yuan et al.,2016) obtained textual information mainly from the project's title, blurb, introduction, as well as comments interaction between creators and investors. One of representative work (Yuan et al.,2016) is to use topic models DC-LDA to extract topical features from project background story, thus achieved good performance both on prediction and interpretation.

CF Prediction with Multimodal. With scalable multimodality dataset, crowdfunding prediction with multimodality often achieves stronger performance than with unimodal and has attracted more research interests. As reported in (Kaminski.,2020), using text, image, and text signal transcribed by audio analysis followed by feature fusion, the success probability can be predicted accurately. However, transcribe audio into text will consume a lot of time and make a loss on audio feature. Chaoran et al. (2019) replaces audio signals with meta data and proposes MDL model, which uses multiple convolutional layers and achieves good classification performance. Similarly, Shi et al. (2021) trained a RNN network, consisting of a pre-trained VGGish to extract MEL features of audio directly and federated metadata to make predictions. In addition to prompting the prediction modal as the combination of text, images, audio and meta, Tang et al. (2022) proposed DCAN, a deep cross-attention network with video modality into the reasoning architecture, achieved SOTA result so far.

Challenge in CF Prediction. The performance of prediction model relies on the quality of modalities. Single modality typically holds the issues of sampling time constraint and single semantic information. For example, mete data often lacks features such as supporters and number of comments that can only be obtained after the event, which significantly limits the use of some of the meta information that play a significant role in forecasting. The text introduction of some projects tends to use imitative storylines and it is difficult to illustrate the traits of the projects, leading to a poor learning for generalization. The most straightforward way (Kaminski.,2020) to mitigate those problems is to add more modalities as the input. For example, it is possible to extract images and textual description of the proposed product and then fed them into image-text tasks. Another trend (Tang et al., 2022) is to join the video into encoder for video understanding tasks regarding to the attractiveness and authenticity of a compelling video.

However, there is several information redundancies in the encoding process. Crowdfunding involves actual interest transactions, which necessarily requires a lot of signals and clues to alleviate information asymmetry. For 75% of the projects presented in the platform, the text introduction is more than 863 words, and the video duration is above 214 seconds (Appendix A2). However, existing dataset for visual-language models trained normally about action recognition and have few seconds for each video such as UCF-101, Sports-1M and Kinetics. In the crowdfunding context, projects in multiple scenarios contain complex semantic information which cannot be solved only by action recognition. Therefore, how to extract many video frames and text sequences concisely and effectively is the first challenge. The second problem is how to align different modalities and represent them into the same semantic space. It is notable that image to text alignment data is relatively simple to obtain on the Internet, but other types of alignment data may be difficult to collect. There is not yet a crowdfunding dataset with fine-grained annotations available for research due to the high cost.

2.2 Multimodal Reasoning

Existing studies have proposed various architectures of multimodality reasoning and can be summarized in three parts: Feature encoder, multimodality fusion as well as a decoder (optional).

About the visual encoding, there are commonly three types. The first type is to use the object detection model (generally Faster R-CNN) to identify the target region in the image and generate the feature representation of each target region for input into the subsequent model (Li et al.,2020). The second way is to use CNN model to extract grid feature as image side input (Tang et al.,2022). Another way is to decompose the image into patches adopted by ViT (Dosovitskiy et al.,2020), and each patch generates embedding and be fed into the model. It is more efficient than the previous two methods and does not need to rely on the object detection module or the pre-CNN feature extraction module.

With transformer shining in language models, text coding basically adopts BERT model (Devlin et al.,2018) and its variants such as RoBERTa (Liu et al.,2019) and DeBERTa (He et al.,2020), which are good at capturing various temporal information.

For the modalities alignment and fusion, there are two major research lines: (i) co-attention. Transformer coding is used for image side and text side respectively, and cross attention between image and text is added in the middle of each Transformer module; (ii) merged attention. the information on the image side and the text side is spliced together at the beginning and input into the Transformer model. As reported in (Lu er al.,2016), using co-attention can capture the complex associations between different modalities. However, this approach also shows a complex calculation degree.

Multimodal reasoning has been widely applied in many fields. In the study by Merler et al(2019)., they used multimodal excitement features for automatic curation of sports highlights, which leverages image, audio, and textual data t o identify the

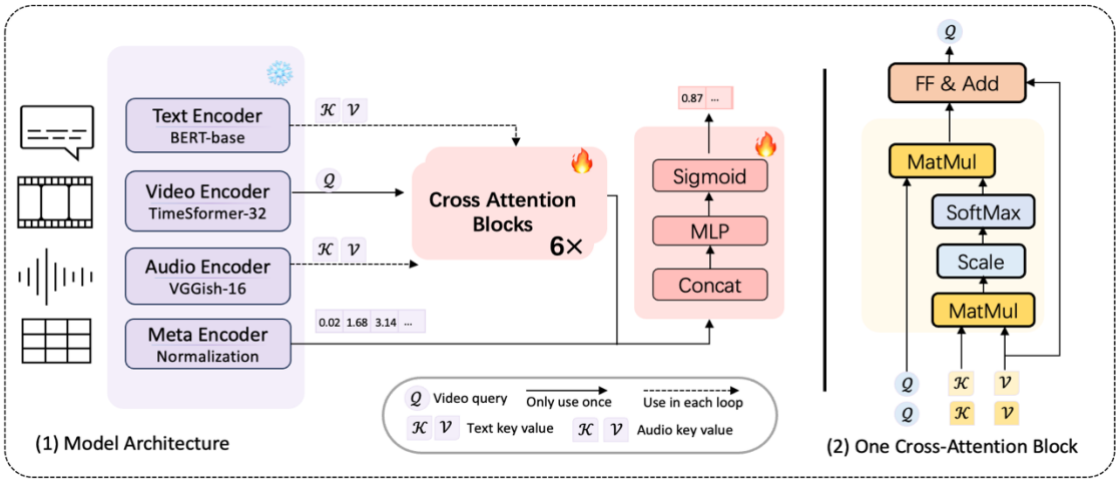


Figure1. (1) The overview of our method. Given M text sequences, N video frames and audio windows, the method use 6 cross attention blocks to align the three modalities into visual semantic space. Then the output of last layer for each modality will be combined with meta representation before being fed into the MLP decoder. **(2) One cross-attention block.** It uses the query from video and key-values pairs from text and audio to generate the next new video query and do the self-attention every round at the same time.

most exciting moments by analyzing key moments of the game. This method provides an effective way to automatically extract exciting moments from a large amount of sports events, offering a better viewing experience for the audience. Zhang et al. (2020)'s research focused on the prediction of Video Quality of Experience (QoE). They proposed a deep learning framework that predicts the user's video viewing experience by analyzing factors such as video visual quality, audio quality, and buffering. This method can help video service providers improve their service quality and increase user satisfaction. Evangelopoulos (2013) used multimodal saliency and fusion methods for movie summarization help viewers understand the content of the movie more quickly, saving viewing time. Video classification is another main battlefield for multimodal applications, and the accuracy and reliability of multimodal applications have been supported by many fruitful works (Y.-G. Jiang et al., 2018; Yang et al.,2016; Zhao,2019).

In conclusion, these studies all demonstrate that multimodal reasoning provides a powerful tool for processing and understanding complex multimedia data.

3 Method

As shown in Figure1. The method has three major procedures: encoding, interaction, and decoding. We feed the audio, text, video and meta into feature encoder to obtain the acoustic representation, textual representation as well as the visual and meta representation. Then, we use the query generated by visual representation and key-value pairs generated by acoustic and textual representation to align and fusion the three modalities. The output of last layer for each modality will be combined with meta representation before being fed into the MLP decoder. The pseudocode provided in Figure 2

Encoding. The model takes four modalities as the input and obtains each representation by the following functions:

$$A = AudioEncoder(X_{audio}) \quad (1)$$

$$T = TextEncoder(X_{text}) \quad (2)$$

$$V = VideoEncoder(X_{video}) \quad (3)$$

$$M = MetaEncoder(X_{meta}) \quad (4)$$

Where AudioEncoder (·) is implemented as a cnn architecture, TextEncoder (·) and VideoEncoder(·) are based on transformer architecture. Specially, we use the Mel spectrum of the audio file, and then input it into the Vggish model (Hershey et al.,2017) to obtain the last layer in the model as audio representation. Meanwhile, we use the hidden states of the last layer in BERT (Devlin et al.,2018) as TextEncoder (·). In order to capture long sequences information in different scenarios, VideoEncoder (·) uses the classification layer of TimeSformer (Bertasius et al.,2021) trained on HowTo100M as the video feature representation. The MetaEncoder is a normalization layer.

Interaction. After obtaining all modal representation, we use two linear layers convert them into keys and values, respectively. For the query, inspired by (Tang et al.,2022), we use the self-attention to generate a query. Now. we have a set of new queries ($\boldsymbol{q}$), key ($\boldsymbol{\mathcal{K}}$) and value ($\boldsymbol{\mathcal{V}}$) for each modal, then we fed them into cross-attention blocks. The video-audio cross attention output $H_{(i)}^{va\_attn}$,video-text cross attention output $H_{(i)}^{vt\_attn}$ and video self-attention output $H_{(i)}^{v\_attn}$ are defined as:

$$H_{(i)}^{va\_attn} = \mathcal{V}_{audio}^{T} \cdot Softmax\left(\frac{\mathcal{K}_{audio} q_{(i)}^{va}}{\sqrt{d_{cross}}}\right) \quad (5)$$

$$H_{(i)}^{vt\_attn} = \mathcal{V}_{text}^{T} \cdot Softmax\left(\frac{\mathcal{K}_{text} q_{(i)}^{vt}}{\sqrt{d_{cross}}}\right) \quad (6)$$

$$H_{(i)}^{v\_attn} = \mathcal{V}_{video}^{T} \cdot Softmax\left(\frac{\mathcal{K}_{video} q_{(i)}^{v}}{\sqrt{d_{cross}}}\right) \quad (7)$$

Where i represents i th cross-attention block, d_{cross} is the dimension in cross-attention blocks. When i=0, $q_{(0)}^{va}, q_{(0)}^{vt}, q_{(0)}^{v}$ all equal the query of video q^{v} . When $i \geq 1$, $q_{(i)}^{va}, q_{(i)}^{vt}, q_{(i)}^{v}$ are the output for i th cross-attention block and are defined as:

$$q_{(i)}^{vt} = FC(H_{(i)}^{vt\_attn}) + q_{(i-1)}^{vt} \quad (8)$$

$$q_{(i)}^{va} = FC(H_{(i)}^{va\_attn}) + q_{(i-1)}^{va} \quad (9)$$

$$q_{(i)}^{v} = FC(H_{(i)}^{v\_attn}) + q_{(i-1)}^{v} \quad (10)$$

Then we fuse these features obtained from the last layer of the cross-attention blocks as well as meta data M. To mitigate

overfitting, we add a dropout layer with a drop rate of 0.3 to each layer. The fused output f is defined as:

$$f = cat(q_{(i)}^{vt}, q_{(i)}^{va}, q_{(i)}^{v}, M). \qquad (11)$$

Algorithm 1
Encoder:
Construct input X=$\{X_{audio}, X_{text}, X_{video}, X_{meta}\}$
input size X_{text} [bs,512] ; X_{video} [bs,32,224,224] ; X_{audio} [bs,128,64,96] ; X_{meta} [bs,4,1]
Extract feature representations of each signals
T [bs,768] ; V [bs,32,600] ; A [bs,128] ; M [bs,4]
Query, key, value generator
Q [bs,512,1] $\mathcal{K}_{audio}$, $\mathcal{V}_{audio}$[bs,1,512] $\mathcal{K}_{text}$, $\mathcal{V}_{video}$ [bs,1,512]
Interaction: while i<= 5 do :
Calculate attention weights, and use them to compute new feature values
$H_{(i)}^{va_{attn}}$, $H_{(i)}^{vt\_attn}$, $H_{(i)}^{vt\_attn}$ [bs,512,1]
Use attention to compute new feature values
##$q_{(i)}^{vt}$, $q_{(i)}^{va}$, $q_{(i)}^{v}$ [bs,512,1]
Concat three feature values as well as meta feature to obtain fuse feature f
##$f = cat(q_{(i)}^{vt}, q_{(i)}^{va}, q_{(i)}^{v}, M, axis = 1)$. **End while**
Decoder:
Feed f to the decoder to obtain the prediction value
y=MLP(f)

Figure2. The pseudocode of our method. Our proposed multimodal network consists of three main components: encoder, interaction, and decoder. The interaction module aims to align and fuse different input modalities.

4 Experiments

4.1. Dataset

We collected a multimodal crowdfunding dataset with 60k video-text pairs [. Limited by computer storage space, we only used half of the data for the experiment, which we called Kick30K. It covers 15 categories,172 sub-categories and 2 statuses on Kickstarter, the most famous crowdfunding platform in the world. Each category with 1,000 positive and 1,000 negative samples, respectively. Each sample contains a project video, text introduction, metadata, and a label for success or failure. The dataset is split into training, validation, and test splits with 18,000, 6,000 and 6,000 samples, respectively.

Besides, Kick60K will be used for extensive experiment to explore the weights of different modality in the module. It contains around 68,000 such samples across 15 categories distributed all the word. 18,000 of them have at least one missing modality. We illustrate more basic sample details of the two datasets in Appendix A1.

The data processing strategy used as following: First, we use 0.95 quantile as cut-off for each modality. For video modality, we sample 32 frames uniformly for each video and crop each frame to 224*224, and then fed them into TimeSformer pre-trained on HowTo100M [1]. TimeSformer stems from ViT and advances to process long video modeling. With divided space-time attention, TimeSformer can learn to focus on the spatial and temporal relevant parts of the video for spatiotemporal understanding. For text modality, we removed the punctuation, stop words and fixed template (risk and challenge notification) from the story text, filtered out words with frequency less than 5 and used BERT-base-uncased tokenizer to tokenize the input [2]. For audio modality, we separated the audio from the video file, calculated the mfcc spectrum at a sampling rate of 128, and converted it into a two-dimensional map of 64*96, then fed into Vggish [3]. In situations where some modalities are not present in the input of the feature backbone, we replace their representations with zero vectors.

Decoding. Finally, the fused output f is fed into an MLP, which has three hidden layers and ReLU activation functions to predict the project. The compete procedure of our model is shown as Algorithm1 in Figure2.

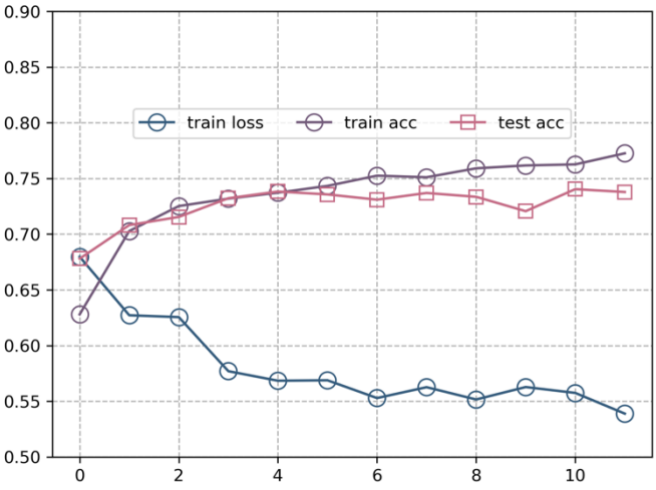
4.2. Implementation

The following part presents the experimental settings of our model and the baseline methods.

Experimental Settings We employ individual feature extractor as crowdfunding datasets differ from general multimodal datasets, which tend to contain rich visual and textual semantic information. Specifically, we adopt 12-layer Transformer's BERT-base as the text extractor [1]. Meanwhile, TimeSformer pre-trained in Dynamics-600 with 32-frame sampling strategy is used as video encoder backbones [2], as well as a VGGish [3] based on VGG18 architecture is used to extract audio features. Also, the encoding of metadata is carried out after normalization.

In model alignment and fusion, our parameters are set as following: cross attention block size is 6 and embedding dimension is 512. In the MLP, the three hidden layer embedding dimension is 768,512,64. In model training, the batch size was set at 4, 30 rounds, and the learning rate was 1e-4. Our experiment was conducted on one NVIDIA GeForce RTX 3090 GPU. Figure3 shows our method ‘s training curve across 12 epochs.

Figure3. Accuracy curve of our method across 12 epochs.



[1]https://huggingface.co/docs/transformers/model_doc/timesformer
[2] https://huggingface.co/bert-base-uncased
[3] https://tfhub.dev/google/vggish/1

Baseline Methods Our benchmark model contains unimodality (Devlin et al., 2018; Kim et al., 2014; Hara et al., 2017; Bertasius et al., 2021) and bimodality (Tang et al., 2022; Shi et al., 2021). Our reproduced results are basically consistent with the original paper, although there is a slight decrease of around 1%. We believe that it is reasonable as we have changed the training device and used a larger dataset. Note that all methods are under the same codebase and same dataset, thus the comparisons are fair. See Appendix B for more details.

Table 1. Comparison with unimodality and bimodality methods on Kick30.

Model	Modality	Size	Game		Theatre		Arts		Dance		Avg	
Metric			Acc.	F1.	Acc.	F1.	Acc.	F1.	Acc.	F1.	Acc.	F1.
TextCNN (Kim,2014)	Text	1M	68.50	76.53	65.15	73.07	67.79	66.45	65.09	69.62	67.46	69.20
RoBERTa (Liu et al.,2019)	Text	1M	73.90	79.23	61.59	61.87	62.32	69.85	72.03	77.57	70.06	73.64
TimSformer (Bertasius et al.,2021)	Video	1M	73.34	82.22	73.22	77.98	65.79	73.18	**73.72**	**80.01**	70.93	76.37
R3D (Hara et al.,2017)	Video	1M	76.58	65.96	62.44	69.43	69.23	73.84	68.67	73.14	71.30	73.52
DCAN (Tang et al.,2022)	Text+Video	27M	80.25	82.32	72.21	69.53	**69.99**	**75.10**	69.05	72.34	71.38	73.99
Basic+ VGGish (Shi et al.,2021)	Meta + Audio	9M	70.09	72.63	69.82	70.53	68.32	71.06	65.33	67.32	68.08	70.41
Ours	Text+Video+Audio+Meta	28M	**81.94**	**85.71**	**78.53**	**82.41**	68.70	74.05	70.22	76.86	**73.25**	**77.99**

4.3. Performance

We perform crowdfunding reasoning tasks with the Kick30 we independently collected. Table1 shows the main results. Column game and theatre are the two results which test on game and theatre category project only, and column arts and dance are the two results which test on arts and dance category project only. The former two achieve top2 accuracy of our method, while the latter are the bottom2.

As shown in Table1, our method outperforms unimodality baseline by 5.79% and surpasses bimodality baseline by performance 1.87%. It is also interesting to see the difference among the 15 categories[4], our method achieves 10/15 performance gain. Table1 also shows 4 category results with the accuracy of top2 and bottom2 by our method.

It is noted that our method extremely good at reasoning in game and theatre category, with an accuracy of 81.94% and 78.53% respectively. Meanwhile, it seems to have no talent for arts and photography. Overall, the results verify the effectiveness of multimodality and the potential of our basic fusion architecture. Figure4 shows the all-category's metric.

5 Analysis

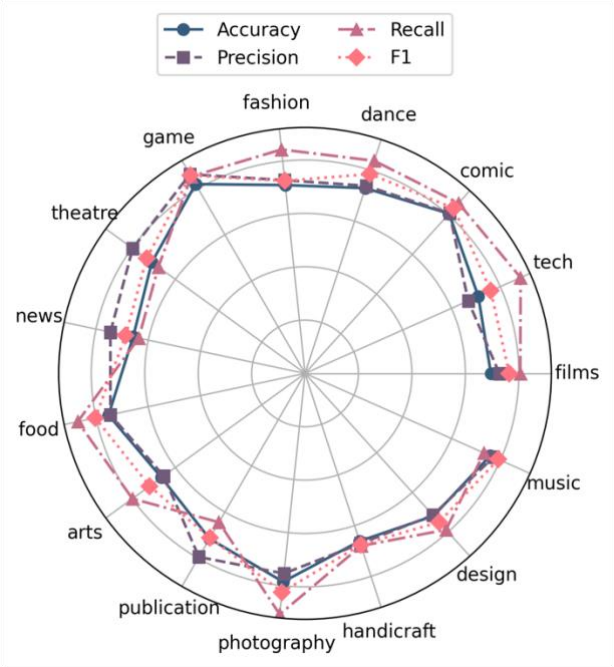
This part aims to investigate why and how our method works and discuss contribution factors as well as limitation.

5.1 Ablation analysis

To further investigate the contribution of each component in our method, we conduct ablation studies in Table2. We first replace each modality's backbone to test its feature extraction ability. However, as we mentioned before, without transformer-based backbones, directly applying convolutional layer tends to capture local semantic information and fails to encode a long sequence, thus result in inferior performance. Next, we compared the difference in semantic space, one is which modality is suitable as the common semantic space, and the other is how deep this space should be. It proves that the video frame space can better align and fuse various modalities. The reason we think is that the video frame covers the whole video space, so it has a strong correlation with the audio as well as the text space provides many tags that can be used as weak supervision for video semantic space. When incorporating all three modalities in the same semantic space, our alignment and fusion network can be stacked deeper for enhance final performance. The results show that the best number of stacking layers is 6.

We also explore the weights of different modality in the module. Table3 shows the results in Kick30 and Kick60. Kick60 dataset contains around 68,000 projects collected form Kickstarter. But 18,000 samples are lack of meta information. We conducted the combined experiments in turn, and reported the average results, but without for each category. Our analysis shows that the classification results is relatively robust with the absence of audio modality, but the performance is greatly reduced when both text and video modality are missing.

Figure4. Comparison results for all categories in Kick30.



[4] More details see in AppendixC1.

Table2. Ablation analysis for different feature backbone and different modalities. The number inside the brackets represents cross attention blocks size

Method	Size	Acc.	F1	Prec.	Recall	Dim
ResNet18	19M	69.74	70.42	79.25	63.36	512
GloVe300	24M	69.28	77.58	66.30	93.49	300
MFCC128	27M	70.45	73.47	75.06	71.95	128
cross_audio blocks(2)	12M	71.52	74.28	77.55	75.88	512
cross_audio blocks(4)	20M	71.69	77.20	71.94	83.30	512
cross_audio blocks(6)	28M	72.03	78.70	73.70	82.16	512
cross_text blocks(2)	12M	71.00	73.35	76.82	70.17	512
cross_text blocks(4)	20M	71.89	75.87	74.10	77.72	512
cross_text blocks(6)	28M	71.46	76.03	72.78	79.57	512

Table 3. Modalities ablation analysis in Kick30 and Kick60.

Modalities				Kick30K		Kick60K	
Meta	Audio	Text	Video	ACC	F1	ACC	F1
✓	✓	✓	✓	**73.25**	**77.99**	-	-
✓	✓	✓	✗	70.79	76.63	72.83	74.99
✓	✓	✗	✗	68.57	72.33	70.30	74.65
✓	✓	✗	✓	70.53	73.62	72.11	72.46
✓	✗	✗	✓	71.38	77.01	73.19	74.05
✓	✗	✓	✓	71.66	77.46	72.52	78.17
✓	✗	✓	✗	72.33	78.82	70.83	70.76
✗	✓	✓	✓	72.90	74.21	**74.36**	**80.28**

Besides, sensitivity analysis studies are conducted for the hyperparameters, dropout rate learning rate. Notably, without dropout layer or a higher learning rate for training, the training curve fails to converge and shows a heavy overfitting. When randomly dropout some features after the layer of activation, such phenomenon was significantly alleviated.

Table4. Hyperparameters analysis.

Modification	Acc.	F1
w/o ReLU	73.02	74.81
w Tanh	72.71	75.47
w Sigmoid	73.16	76.57
w/o dropout	-	-
w dropout (0.3)	73.25	77.99
w dropout (0.5)	73.07	78.46
Lr(5e-4)	72.96	76.83
Lr(1e-4)	73.25	77.99
Lr(1e-3)	-	-

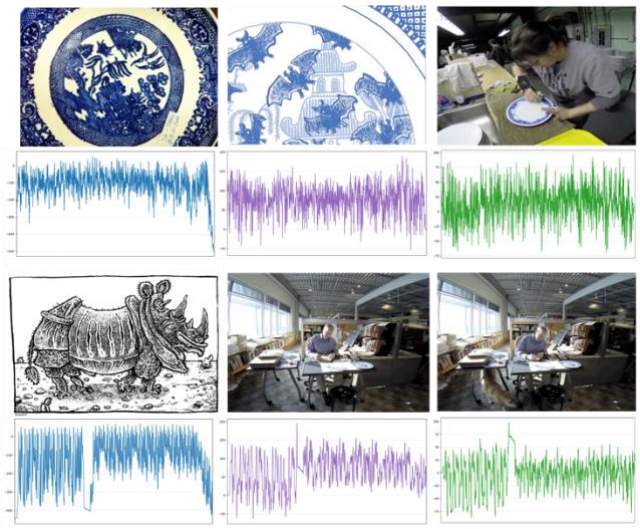


Figure5. Visualization analysis.

5.2 Visualization analysis

Figure 5 shows the modality visual comparison between video frame and audio with top3 and bottom3 attention weight each. This project is about the design of antique plates. We observed that high-weight video frames more clearly reflected product visual details and craftsmen working scenes.

5.3 Limitation analysis

In our study we examined the benefit of our framework only for classification tasks with 2 weak labels. As it is the most direct way to consider the results of the crowdfunding project as the supervised label. Future work can still follow the supervised route, with adding more labels as supervision. We also try to further subdivide the classification label according to the proportion of the actual amount of financing to the target amount and regard it as a multiple classification task. But the performance is not better than the present way, which demonstrate our framework still lacks deep reasoning skills. Another path is self-supervised, such as CLIP (Radford et al.,2021) and ALIGN (Jia et al.,2021). Such approaches have impressive ‘zero-shot’ performance because the information inside the crowdfunding dataset is richer than the annotated one. Besides, it can also leverage distillation methods to do very large multi-modality training to obtain robust results.

A further shortcoming of this work is that we only included the dataset from Kickstarter website, although it involves 15 categories and 172 sub-categories. However, it is still difficult to eliminate the noise caused by the choice of platform, because the form and credibility of information presented by each platform are different. Therefore, it is necessary to conduct further validation of our approach on other popular websites (Indiegogo, GoFundMe, JustGiving, et al.).

5 Reference

Yang Z, Fang Y, Zhu C, et al. i-code: An integrative and composable multimodal learning framework[C] *Proceedings of the AAAI Conference on Artificial Intelligence.* 2023, 37(9): 10880-10890.

Liang, T. P., Wu, S. P. J., & Huang, C. C. (2019). Why funders invest in crowdfunding projects: Role of trust from the dual-process perspective. *Information & Management*, 56(1), 70-84.

Hershey S, Chaudhuri S, Ellis D P W, et al. CNN architectures for large-scale audio classification[C] *2017 IEEE international conference on acoustics, speech, and signal processing (icassp)*. IEEE, 2017: 131-135.

Girdhar R, El-Nouby A, Liu Z, et al. Imagebind: One embedding space to bind them all[C] *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition.* 2023: 15180-15190.

Lu J, Yang J, Batra D, et al. Hierarchical question-image co-attention for visual question answering[J]. *Advances in neural information processing systems*, 2016, 29.

Yuan H, Lau RYK, Xu W. The determinants of crowdfunding success: A semantic text analytics approach[J]. *Decision Support Systems*, 2016, 91: 67-76.

Tang Z, Yang Y, Li W, et al. Deep cross-attention network for crowdfunding success prediction[J]. *IEEE Transactions on Multimedia*, 2022, 25: 1306-1319.

Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of deep bidirectional transformers for language understanding[J]. arXiv preprint arXiv:1810.04805, 2018.

M. Merler et al., Automatic curation of sports highlights using multimodal excitement features, *IEEE Transactions on Multimedia*, vol. 21, no. 5, pp. 1147–1160, May 2019.

H. Zhang et al., DeepQoE: A multimodal learning framework for video quality of experience (QoE) prediction, *IEEE Transactions on Multimedia*, vol. 22, no. 12, pp. 3210–3223, Dec. 2020.

G. Evangelopoulos et al., Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention,"*IEEE Transactions on Multimedia*, vol. 15, no. 7, pp. 1553–1568, Nov. 2013.

Y.-G. Jiang et al., Modeling multimodal clues in a hybrid deep learning framework for video classification, *IEEE Transactions on Multimedia*, vol. 20, no. 11, pp. 3137–3147, Nov. 2018.

Dosovitskiy A, Beyer L, Kolesnikov A, et al. An image is worth 16x16 words: Transformers for image recognition at scale[J]. arXiv preprint arXiv:2010.11929, 2020.

Li G, Duan N, Fang Y, et al. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training[C]. *Proceedings of the AAAI conference on artificial intelligence.* 2020, 34(07): 11336-11344.

Kaminski J C, Hopp C. Predicting outcomes in crowdfunding campaigns with textual, visual, and linguistic signals[J]. *Small Business Economics*, 2020, 55: 627-649.

Cheng C, Tan F, Hou X, et al. Success Prediction on Crowdfunding with Multimodal Deep Learning[C] *IJCAI. 2019*: 2158-2164.

Lee S H, Shafqat W, Kim H. Backers Beware: Characteristics and Detection of Fraudulent Crowdfunding Campaigns[J]. *Sensors*, 2022, 22(19): 7677.

Bertasius G, Wang H, Torresani L. Is space-time attention all you need for video understanding? [C] *ICML*. 2021, 2(3): 4.

Hara K, Kataoka H, Satoh Y. Learning spatial-temporal features with 3d residual networks for action recognition[C] Proceedings *of the IEEE international conference on computer vision workshops*. 2017: 3154-3160.

Zhu, Tong, et al. Multimodal Emotion Classification with Multi-level Semantic Reasoning Network. *IEEE Transactions on Multimedia* (2022).

Yang, Xiaodong, Pavlo Molchanov, and Jan Kautz. Multilayer and multimodal fusion of deep neural networks for video classification. *Proceedings of the 24th ACM international conference on Multimedia.* 2016.

Zhao, Tianqi. Deep multimodal learning: An effective method for video classification. *2019 IEEE International Conference on Web Services (ICWS)*. IEEE, 2019.

Kim Y. Convolutional neural networks for sentence classification[J]. arXiv preprint arXiv:1408.5882, 2014.

Wei, Lingwei, et al. Modeling Both Intra-and Inter-Modality Uncertainty for Multimodal Fake News Detection. *IEEE Transactions on Multimedia* (2023).

Shi J, Yang K, Xu W, et al. Leveraging deep learning with audio analytics to predict the success of crowdfunding projects[J]. *The Journal of Supercomputing*, 2021, 77: 7833-7853.

Liu, Fan, et al. Disentangled multimodal representation learning for recommendation. *IEEE Transactions on Multimedia* (2022).

He P, Liu X, Gao J, et al. DeBERTa: Decoding-enhanced Bert with disentangled attention[J]. arXiv preprint arXiv:2006.03654, 2020.

Liu Y, Ott M, Goyal N, et al. Roberta: A robustly optimized Bert pretraining approach[J]. arXiv preprint arXiv:1907.11692, 2019.

Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C] International conference on machine learning. *PMLR*, 2021: 8748-8763.

Jia C, Yang Y, Xia Y, et al. Scaling up visual and vision-language representation learning with noisy text supervision[C] International conference on machine learning. *PMLR*, 2021: 4904-4916.

6 Appendix

A. Dataset

A1. Data Sample

This section presents a data sample which collected from Kickstarter. It contains 3 part shown in Figure I, meta [project status, project goal, project category, project launch time and end time, project amount of creator], text [A background story without images about the project], video[An introductory video with sound normally lasts about two minutes]. Also, Table A and Table B present the split detail of the two dataset and the dataset structure.

Kick30 contains 30,000 such samples across 15 categories distributed 52 states in the United States. Each sample contains a complete text-video pair. These projects were initiated from April 2011 to September 2023.

Kick60 contains around 60,000 such samples across 15 categories distributed all the word. Some samples have at least one missing modality. These projects were initiated from March 2010 to September 2023.

Table A. The split detail of datasets for Kick30 and Kick60.

Dataset	Category	Train	Val	Test
Kick30	film	1.2k	0.4k	0.4k
	tech	1.2k	0.4k	0.4k
	comic	1.2k	0.4k	0.4k
	dance	1.2k	0.4k	0.4k
	fashion	1.2k	0.4k	0.4k
	game	1.2k	0.4k	0.4k
	theatre	1.2k	0.4k	0.4k
	news	1.2k	0.4k	0.4k
	food	1.2k	0.4k	0.4k
	arts	1.2k	0.4k	0.4k
	publication	1.2k	0.4k	0.4k
	photography	1.2k	0.4k	0.4k
	handicraft	1.2k	0.4k	0.4k
	design	1.2k	0.4k	0.4k
	music	1.2k	0.4k	0.4k
	total	18k	6k	6k

Dataset	Category	Train	Val	Test
Kick60	film	2.5k	0.8k	0.8k
	tech	3k	1k	1k
	comic	2.9k	1k	1k
	dance	2.2k	0.7k	0.7k
	fashion	1.8k	0.6k	0.6k
	game	2.7k	0.9k	0.9k
	theatre	2.5k	0.8k	0.8k
	news	4.2k	1.4k	1.4k
	food	2.1k	0.7k	0.7k
	arts	1.7k	0.6k	0.6k
	publication	1.9k	0.6k	0.6k
	photography	1.7k	0.6k	0.6k
	handicraft	3k	1k	1k
	design	2.3k	0.8k	0.8k
	music	2.5k	0.8k	0.8k
	total	61k	12k	12k

Table B. Data structure for Kick30 and Kick60.

Modality	Feature	description
Visual	Video Introduction	A video clips about the project normally over a minute
Textual	Background Story	A text paragraph usually contains the details of the project
Acoustic	Background Audio	A continuous audio frame extracted from the video
Structured	Launch_date	The project launch date, json format,yy/mm/dd
	Finish_date	The project finish date, json format, yy/mm/dd
	Duration	The project last days, json file
	Goal	The project fundraising goal, also is the threshold whether this project can obtain the all raised money.
	Finish_amount	The project actual raised money.
	Result	IF finish>=goal, result=1;else,result=0.
	Create_amount	The number of previous projects initiated by the creator
	Category	The project type defined of Kickstarter

A2. Data Statistics

Here we present more detailed statistics around our dataset. Figure1 presents the video introduction time distribution (seconds) and Figure2 presents the video introduction frames distribution. Figure3 shows the distribution of text words in story background. Figure4 illustrates the launch date(year) and Figure5 shows the duration of projects. Figure6 presents the average of video time and frames of all categories.

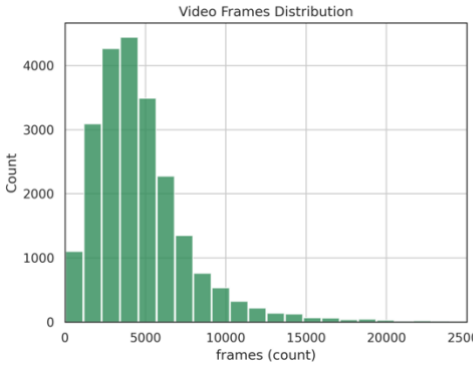


Figure1.

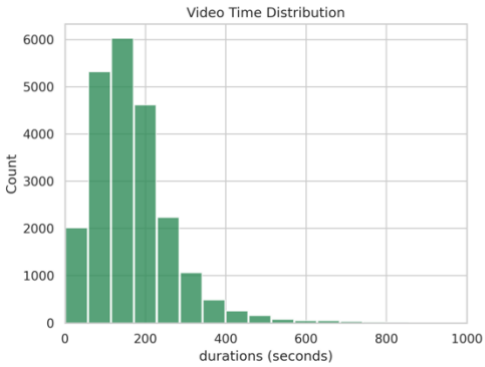


Figure2.

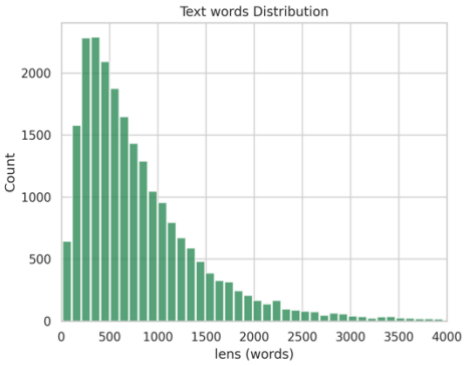


Figure3.

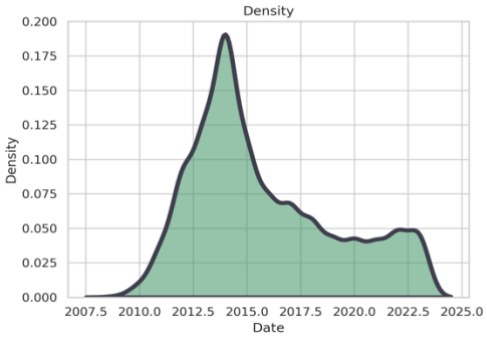


Figure4.

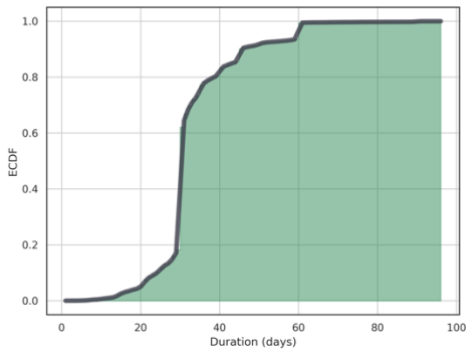


Figure5.

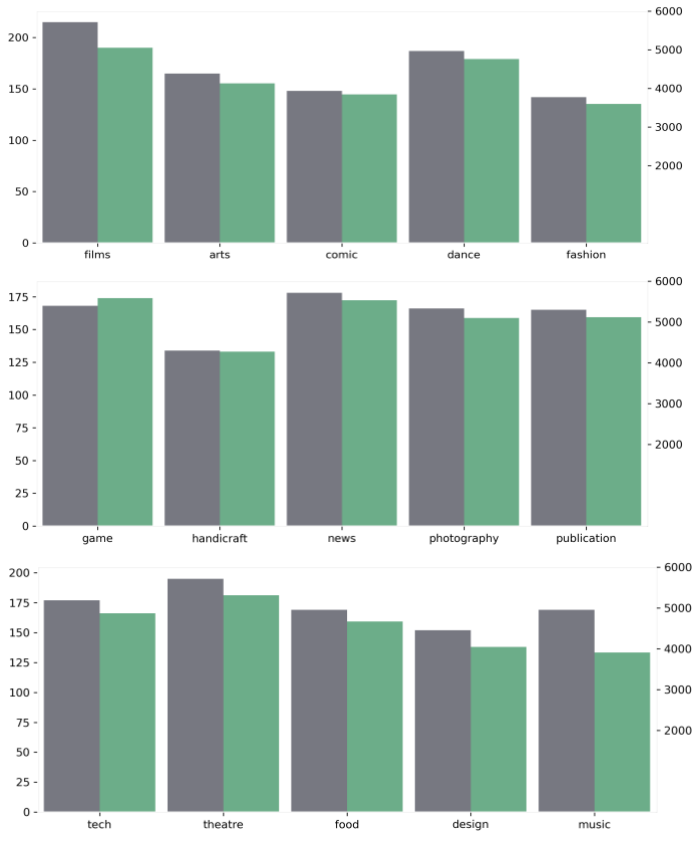


Figure6.

B. Baseline Setup

Our baselines include two lines: unimodality and bimodality.

(1) For all unimodality methods, we adopt fine-tuning backbone plus a MLP with three hidden layers. Specifically, the TextCNN (Kim,2014) backbone uses three convolution kernels, each with a size of 3, 4, and 5, and the number of channels is 100. RoBERTa (Liu et al., 2019) and TimSformer (Bertasius et al., 2021) used the setup in the original article. The number of convolution layers in R3D (Hara et al., 2017) is 10, and the size of convolution kernel is 3*3*3.

(2) For bimodality methods, DCAN (Tang et al.,2022) and Basic+ VGGish (Shi et al.,2021) also used the same setup in the original article.

C. Performance for all categories of crowdfunding project.

Model	Modal	Size	Film		Arts		Comic		Dance		Fashion		Game	
Metric			Acc.	F1.	Acc.	F1.	Acc.	F1.	Acc.	F1.	Acc.	F1.	Acc.	F1.
TextCNN	Text	1M	65.93	69.39	67.79	66.45	68.19	67.77	65.09	69.62	66.16	61.23	65.09	70.06
RoBERTa	Text	1M	68.55	75.43	62.32	69.85	76.71	81.14	72.03	77.57	68.64	69.44	69.32	74.33
TimSformer	Video	1M	68.96	76.02	65.79	73.18	75.43	80.68	73.72	80.01	69.26	68.43	73.34	85.43
R3D	Video	1M	69.25	76.50	69.23	73.84	78.29	76.32	68.67	73.14	59.14	63.33	76.58	71.22
DCAN(Zhe et al.,2022)	Text+Video	27M	69.43	65.21	69.99	75.10	79.33	82.34	69.05	72.34	68.49	69.17	80.25	82.32
Basic+ VGGish	Meta + Audio	9M	65.23	67.77	68.32	71.06	67.98	75.32	65.33	67.32	69.83	70.21	70.09	72.63
Ours	Text+Video+ Audio+Meta	28M	70.33	79.20	68.70	74.05	80.57	83.29	70.22	76.86	70.99	72.58	81.94	85.71

		Size	Photography		Publication		Tech		Theatre		Food		Design	
TextCNN	Text	1M	66.58	68.58	67.51	67.29	67.51	66.53	69.12	71.36	68.43	72.22	68.08	72.44
RoBERTa	Text	1M	74.18	79.23	65.43	70.90	70.52	70.17	74.14	79.36	65.10	66.53	70.87	69.92
TimSformer	Video	1M	69.50	77.53	66.15	74.07	74.76	78.06	71.32	78.32	69.85	69.14	67.81	74.00
R3D	Video	1M	74.90	80.23	62.59	62.87	74.34	75.74	76.40	80.48	68.36	63.31	69.34	72.57
DCAN(Zhe et al.,2022)	Text+Video	27M	70.40	83.22	74.22	78.98	70.09	73.44	72.21	69.53	70.33	72.89	71.93	73.26
Basic+ VGGish	Meta + Audio	9M	67.32	66.96	63.44	70.43	69.47	71.75	69.82	70.53	69.98	70.37	68.42	70.31
Ours	Text+Video+ Audio+Meta	28M	71.94	78.21	68.23	74.42	71.52	72.67	78.53	82.41	70.91	73.56	72.78	77.00

		Size	Handicraft		News		Music		Avg	
TextCNN	Text	1M	66.09	68.87	67.84	66.15	71.22	72.02	67.46	69.2
RoBERTa	Text	1M	69.88	72.31	64.78	67.83	69.3	71.11	70.06	73.64
TimSformer	Video	1M	69.53	75.80	71.19	78.69	68.77	68.3	70.93	76.37
R3D	Video	1M	70.93	74.90	72.22	77.73	65.33	70.48	71.3	73.52
DCAN(Zhe et al.,2022)	Text+Video	27M	70.94	70.97	64.27	68.78	69.83	72.32	71.38	73.99
Basic+ VGGish	Meta + Audio	9M	68.96	72.32	64.44	65.85	72.58	73.33	68.08	70.41
Ours	Text+Video+ Audio+Meta	28M	72.63	77.59	74.87	79.65	72.33	77.24	73.25	77.99