Multimodal Intent Discovery from Livestream Videos

Adyasha Maharana[1], **Quan Tran**[2], **Franck Dernoncourt**[2], **Seunghyun Yoon**[2]
Trung Bui[2], **Walter Chang**[2], **Mohit Bansal**[1]
[1]University of North Carolina at Chapel Hill [2]Adobe Research
{adyasha, mbansal}@cs.unc.edu, {qtran, dernonco, syoon, bui, wachang}@adobe.com

Abstract

Individuals, educational institutions, and businesses are prolific at generating instructional video content such as "how-to" and tutorial guides. While significant progress has been made in basic video understanding tasks, identifying procedural intent within these instructional videos is a challenging and important task that remains unexplored but essential to video summarization, search, and recommendations. This paper introduces the problem of instructional intent identification and extraction from software instructional livestreams. We construct and present a new multimodal dataset consisting of software instructional livestreams and containing manual annotations for both detailed and abstract procedural intent that enable training and evaluation of joint video and text understanding models. We then introduce a multimodal cascaded cross-attention model to efficiently combine the weaker and noisier video signal with the more discriminative text signal. Our experiments show that our proposed model brings significant gains compared to strong baselines, including large-scale pretrained multimodal models. Our analysis further identifies that the task benefits from spatial as well as motion features extracted from videos, and provides insight on how the video signal is preferentially used for intent discovery. We also show that current models struggle to comprehend the nature of abstract intents, revealing important gaps in multimodal understanding and paving the way for future work.[1]

1 Introduction

Instructional videos have become increasingly ubiquitous as users generate diverse "how-to", DIY, and tutorial videos. A Pew Research Center 2018 survey of U.S. adult YouTube users (Smith et al., 2018) found that over half of surveyed users use video content to learn how to do things they had not done before. These instructional videos convey both abstract and specific intent for physical tasks such as cooking where e.g., an abstract culinary intent is "let's bring out the flavor" and a detailed intent is "add a pinch of nutmeg". Thus, a key task in instructional video understanding is to discover both abstract and detailed intents. By discovering these intents, we can enable or improve important tasks such as semantic indexing of videos (Kofler et al., 2016), knowledge graph creation for video search and recommendations (Pei et al., 2011; Kofler et al., 2014), intent highlighting, and video summarization (Nalla et al., 2020).

An important domain with rich and complex examples of both abstract and detailed intent types are software training videos for creative tasks such as making photo or video effects. These types of software training videos have been shown to be effective for enhanced learning (Van der Meij, 2017) and are also considered a valuable resource in the era of online learning (Meyer, 2015). Existing video and phrase datasets such as HowTo100M (Miech et al., 2019) cover a wide variety of tutorials for visual tasks demonstrated by humans; however, software-based instructional videos are not a part of such corpora. Hence, in this paper, we present a new corpus of software-instructional videos containing instructional intents, which are derived from Behance Livestreams demonstrating the use of Adobe Photoshop software.[2]

Intent detection has been well-studied in dialogue systems (Wu et al., 2020), but is less explored for instructional video content, especially emerging livestream content (Fraser et al., 2019). While rich in complex procedural instruction and intent, the interactive and social nature of livestreams poses unique challenges. Analyzing language features alone will provide only limited information about

[1]Code and data are available at https://github.com/adymaharana/VideoIntentDiscovery.

[2]https://www.behance.net/live, https://www.adobe.com/products/photoshop.html

Findings of the Association for Computational Linguistics: NAACL 2022, pages 476 - 489

the actual instructional intent and the tools and commands used. For instance, the phrase "flipping the canvas" in "Are you flipping the canvas?" indicates a tool intent, but a closer look at the video clip reveals that it is in fact part of livestream chit-chat and does not take place on-screen. Incorporating both language and video modalities can enhance intent extraction of such ambiguous intents. Hence, in this paper, we present a new joint language-video intent discovery task and a multimodal dataset consisting of: Behance Intent Discovery, and the Behance Livestream video and transcript corpus that intents are found in. We frame intent discovery as a sequence labelling task; each sample in the intent discovery dataset contains a transcribed phrase annotated with token-level tags for abstract and detailed intents, and an associated video timestamp. Our goal is to predict the instructional intents from the transcript in each video.

To perform intent discovery within instructional videos, we propose a multimodal cascaded cross-attention model to predict both the abstract and detailed procedural intents that are present. Additionally, we use late fusion of multimodal embeddings to prevent the visual modality from overwhelming the textual signal, and show significant improvements on the video-based intent detection task using unimodal and multimodal pretrained models like HERO (Li et al., 2020). Further, we compare the performance of various video feature extractors as well as different video lengths, and present benchmark results on the proposed dataset. We find that discovery of tool intents benefit from sparsely-sampled spatial features while creative intents benefit from densely-sampled motion features. In the absence of motion features, most models struggle to utilize the video signal for identification of creative intents. Further, visualization of cross-attention and visual gate modules in the late fusion model suggests strong and meaningful interaction between the two modalities. Our contributions are:

- We introduce and explore the novel task of video-based multimodal intent discovery, and present an annotated dataset consisting of nearly 20K sentences from 66 livestreams for extraction of procedural intents from instructional videos.
- We release a large corpus of software-based instructional videos (2,049 sessions, 3,128 hours total), accompanied by timestamped transcripts, that can be used for pretraining multimodal models.
- We propose the multimodal cascaded cross-attention model and demonstrate the effectiveness of late fusion of multimodal embeddings in this task.
- We present empirical results for the proposed dataset using unimodal and multimodal approaches, and provide insights from analysis of modelling choices for future research.

2 Related Work

Intent discovery has been widely studied in the context of dialog modelling and generation wherein it has been framed as a binary or multi-class classification problem. The SNIPS (Coucke et al., 2018) and ATIS (Dahl et al., 1994) datasets consist of concise single-sentence texts containing intents with constrained vocabulary and attributes. Several works have explored intent classification of internet posts in the context of racial/radicalized intent (Agarwal and Sureka, 2016), purchase intents (Gupta et al., 2014; Wang et al., 2015), discussion forums (Chen et al., 2013) and health queries (Cai et al., 2017). Vedula et al. (2019) propose open intent discovery with unconstrained vocabulary as a sequence tagging task. Using this framework, we present our dataset on instructional intents.

In the wake of exploding visual social-media content, several image-based multi-modal intent datasets have been previously proposed. Kiela et al. (2020); Aprosio et al. (2020) study abusive language and hateful intent in memes and photo posts. Jia et al. (2021) explore intent categories derived from social psychology and use object localization to integrate visual context in task models. Instagram posts are another interesting source for multimodal content (Chen and Hsieh, 2020; Kruk et al., 2019). We introduce the task of video-based multimodal intent discovery, which has been unexplored.

Several tasks have been proposed in the recent years to probe joint video and text understanding. Lei et al. (2018); Kim et al. (2017); Maharaj et al. (2017); Jang et al. (2017); Tapaswi et al. (2016) and Yi et al. (2020) introduce video-based question answering datasets created from various sources of creative visual content, i.e. movies, TV shows, GIFs etc. Lei et al. (2020b) and Lei et al. (2020a) propose the task of video-moment retrieval and next frame prediction respectively, based on query subtitles, while Liu et al. (2020) present the multimodal version of natural language inference. Early

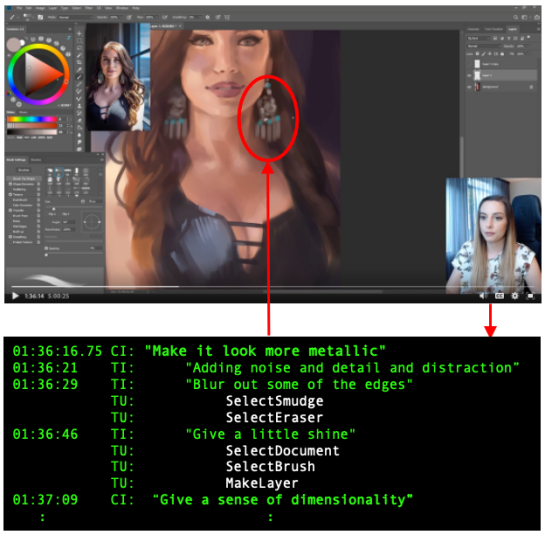


Figure 1: Examples of the output predictions for candidate creative and tool intents given an instructional livestream video and its associated transcript.

models for performing these tasks involve combining pretrained image representations from sparsely sampled videos, and text encodings from pretrained encoders (Devlin et al., 2019; Liu et al., 2019) in architectures for modelling global-local interactions (Zhu and Yang, 2020; Yang et al., 2020), temporal localization (Kim et al., 2019; Zhang et al., 2020), graph-based reasoning (Huang et al., 2020) etc. More recent attempts involve pretraining models on large video+text corpora (Miech et al., 2019) and finetuning on downstream tasks (Sun et al., 2019; Cho et al., 2021; Tang et al., 2021; Lei et al., 2021; Luo et al., 2020). We explore late-fusion of video and text embeddings (Yu et al., 2020) for intent detection in pretrained and non-pretrained multimodal settings.

3 Problem Setting: Intent Discovery from Livestreams

In our setting, each video captures a Behance livestream in which an instructor demonstrates the steps needed to accomplish various image editing or compositing tasks. We specifically focus on rich creative instructional or tutorial livestreams that teach photo editing and compositing methods using an image application such as Photoshop, which consists of over 1,300 basic menu commands and subcommands, tool icons, panels, and galleries.

The livestream itself consists of a screencast of the instructor's application software, a smaller video window showing the instructor, a time-coded transcript of the dialog within the session, and a

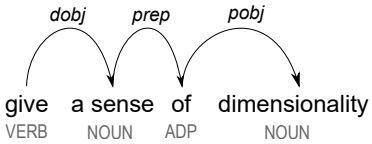


Figure 2: Example of dependency structure of an intent.

Attribute	Statistics
#Sessions	3,356
Min/Max/Avg. session length	1/426/80 mins.
Min/Max/Avg. phrase length	1/142/8 words
Min/Max/Avg. #phrases per session	1/4,552/587
#Distinct tools in corpus	282

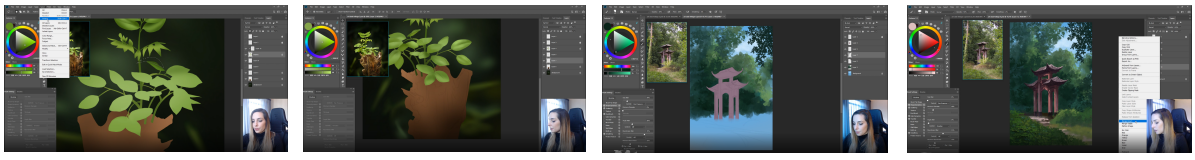
Table 1: Statistics of the Behance Livestreams corpus for sessions and transcribed phrases.

tool timeline which is a time-coded log of the specific application tools used during the livestream. Given the transcript of an instructional video, the video itself, and optional Tool Usage (TU) information from the tool timeline, Fig. 1 shows examples of the Creative Intents (CI, shown at 01:36.16 and 01:37:09) and Tool Intents (TI, shown at 01:36:21, 01:36:29 and 01:36:46) we seek to discover. Further, we wish to combine joint language and video knowledge to gain improvements in detecting candidate intents that are false positives such as the text at 01:37:09, which is only a parenthetical comment by the instructor.

4 Behance Datasets

Dataset Collection. We first obtain 2,049 videos along with their transcripts and tool timelines from the Behance platform. The tool timeline contains a time-stamped record of the tools used in the software during the tutorial. The average session length is 80 minutes with an average of 587 transcribed phrases per session (see Table 1). The tool timelines contain 282 distinct tools with varying frequencies; `Color`, `Select Brush`, `Select Layer` are some of the most frequent ones. The instructional software-based domain of this dataset is significantly different from existing large corpora drawn from YouTube instructional videos (Miech et al., 2019) and TV content (Lei et al., 2018, 2020b), but it is an important learning resource. Hence, we include the unlabelled **Behance Livestreams** corpus as an addition to the pool of video+text corpora that can be leveraged for continued pretraining of multimodal models and finetuning on downstream tasks relevant to software-based livestream videos.

In order to prepare the intent discovery dataset,



1:34:28.55 --> 1:34:33.45
Select the inverse and **give a little bit of detail** right to that edge.

1:48:14.45 --> 1:48:21.65
I am going to **make a new layer** and **use a clipping mask** again on the wood and I might **use a textured brush.**

3:40:48.59 --> 3:50:51.76
So I am **blurring some of my groups**.

4:47:40.83 --> 4:47:48.12
I'm happy with that, so I **merge it down** and then we'll **make a new layer** and let's **add some more plants and stuff**.

Figure 3: Examples of tool intent (blue) and creative intents (magenta) from the Behance Intent Discovery dataset.

Attribute	Training/Validation/Test
#Sentences	13989/2,105/3,917
#Tool intents	3,478/414/825
#Creative intents	674/106/189
#Livestream videos	54/6/6

Table 2: Statistics of the various splits in the Behance Intent Discovery dataset.

Top Unique Verbs in Action	
Tool Intents	merge, select, add, use, paint, duplicate, make, delete, painting, do, flip, decrease, using, lower, figure, erase
Creative Intents	add, make, give, change, paint, create, convey, fill, animate, use, have
Top Unique Nouns in Object	
Tool Intents	layer, color, mask, things, brush, shapes, selection, tool, shift, opacity, canvas, adjustment, stuff, thing
Creative Intents	colors, details, light, shadow, texture, contrast, highlights, vibe, depth, sense, bounce, feeling, elements

Table 3: Unique words in the intent discovery dataset.

we extract candidate intent phrases from the transcripts of the Behance Livestream corpus. Following Vedula et al. (2019), we define an intent as a text phrase consisting of: (i) an *action* word or phrase, which constitutes a definite task, goal or activity and (ii) an *object*, which represents those words or phrases that the action is going to act or operate upon. We generate the dependency graph of sentences, and extract the VERB node as *action* and the direct object of the VERB as the *object*, along with all other children nodes (see example in Fig. 2).[3] Through manual analysis, we identified two major categories of meaningful intent: tool and creative. *Tool intents* are low-level intents that can be typically mapped to a single tool in the software. *Creative intents* are abstract intents used to describe a high-level creative goal that consist of a complex set of actions or tool intents. For instance, in Fig. 3, "make a new layer" is a tool intent that can be mapped to the tool `Create Layer`, while "add more plants and stuff" is a creative intent. All other intents in the corpus, predominantly from chit-chat statements, are irrelevant to our task. We frame the task of intent discovery as a sequence-tagging problem and tag the intent phrases within each sentence with IOB (inside, outside, beginning) span annotations for the two classes: tool and creative intents. Each sample consists of a timestamped sentence with span annotations and the video session it is extracted from.

Based on the above defined framework, we collect manual annotations of tool and creative intents. We employed two annotators using the UpWork crowdsourcing platform and trained them for identifying intents from Behance videos and transcripts.[4] They were instructed to annotate spans for tool and creative intents within each sentence. The annotations were created using the open-source Doccano annotation tool.[5] In total, we collected annotations for 20,011 sentences from 66 Behance videos. The resulting **Behance Intent Discovery** dataset contains 13,989/2,105/3,917 samples in training, development and test splits respectively (see statistics in Table 2). We do not specify the duration of video clips for each sentence in the annotation. In our experiments, we explore varying clip duration and empirically choose a window of 10 seconds ($\pm$5s) around the sentence's timestamp (see Sec. 8.2). The full video sessions are released for further research.

Dataset Analysis. We analyzed tool and creative intents to find the most frequent, unique verbs and nouns mentioned in the phrases. While there are action verbs which are distinctly tool-specific, such as merge, select, and duplicate, there are many verbs

[3] https://spacy.io/api/dependencyparser

[4] https://www.upwork.com/. Annotators were compensated per the $20/hr rate.

[5] https://github.com/doccano/doccano

which are common to both tool and creative intents such as add, make, and paint. Hence, the task model needs to learn the difference between tool and creative intents to be able to classify intents with similar action verbs into the correct categories. Further, we examined the unique nouns occurring in the intents and found lesser overlap between the two intent classes. Creative intents contain abstract and subjective visual concepts which pose a unique and interesting challenge to multimodal models. See Appendix for probing experiments.

5 Methods for Intent Discovery

Intuitively, as in a lot of instructional sources like text books, the text or audio serves as the primary mode of high-level information transfer, while the video/image signal provides detailed context or demonstration. Thus, we start our exploration using text models, which are built on two pretrained models: RoBERTa (Liu et al., 2019) and GPT2 (Radford et al., 2019). There are several previous works focusing on a limited set of intents (Xia et al., 2018), and thus, treat the problem of intent discovery as a classification problem. In our case, given the vast possibilities of potential intents in our sources, we cast the problem as a span detection problem, and design our models accordingly.

5.1 Unimodal Sequence Labelling

Our text models are designed similar to Named-Entity-Recognition models with a pretrained embedding layer and a sequence classification layer on top. Each phrase in the transcript is annotated separately in the intent dataset, leading to efficient processing. Although it is possible to process longer spans of text, in our annotations, we found out that each sentence usually gives enough information to extract the intent inside it, and extra context (neighboring sentences) does not significantly help the decision. We denote an input sentence as $X = [x_0, ..., x_N]$ with N as the length of the input sentence, $Z = [z_0, ..., z_N]$ denotes the common IOB tags of two classes: creative intent and tool intent. Using text encoder f_{enc}, we extract text encodings E i.e. $E = f_{enc}(X)$. The encodings are then passed to the classifier layer for computing tag probabilities i.e. $\hat{Z} = softmax(W_c * E + b_c)$ where W_c, b_c are parameters of the classifier layer. The model is trained end-to-end using cross-entropy loss i.e. $\mathcal{L}_\theta = -\frac{1}{N}\sum_{i=1}^{N} z_i log(\hat{z}_i)$, where θ represents parameters of the entire model.

5.2 Multimodal Sequence Labelling: Naïve Fusion

Seeking to leverage the video information, in our first attempt, we tried a simple feature fusion between the text signal and the video signal in the sequence labelling framework. We add a cross-attention layer on top of the pretrained text encoder in this naïve joint video-text model and use the output of the cross-attention layer for sequence label classification. Let's denote the video features as V. Our model (see Fig. 4 (a)) is described as follows:

$$\hat{Z} = softmax(W_c * f_{self}(f_{cross}(E, V)) + b_c)$$

where E, f_{self} and f_{cross} are text encodings, self-attention and cross-attention layers respectively. This naïve fusion model, however, does not provide any significant improvements compared to text-only baselines (see Sec. 7). Analysis of the results revealed that the textual features dominate the final decision, especially in the creative intent classes. To understand this behaviour, we performed a pilot task in which a human annotator looks through the video segments and tries to guess the intent without any transcript or audio. Our annotator found the task very difficult, and only possible after watching a very long context window, which partially explains the low performance of this model. The video signal is much more ambiguous than the text signal, and when presented with two sources where one is vastly less informative than the other, the model learns to rely only on the text, leading to no improvement compared to the text-only baseline. Joining two sources of features with different predictive utility is difficult. Given the fact that the video feature extractor is not trained on similar data, the video feature might not contain enough information for a direct intent detection task. Fortunately, our pilot task also reveals an important insight, i.e., the video signal *is good at identifying whether an intent is present or not*. Many intent candidates identified by the text models are not creative or tool intents, but are chitchat utterances from the instructor interacting with the audience. In these cases, we posit that the inactivity presented in the video signal is a strong indication that a creative/tool intent does not occur at the current time window. Using this idea, we propose a cascaded model with deeper interaction between video and text signal.

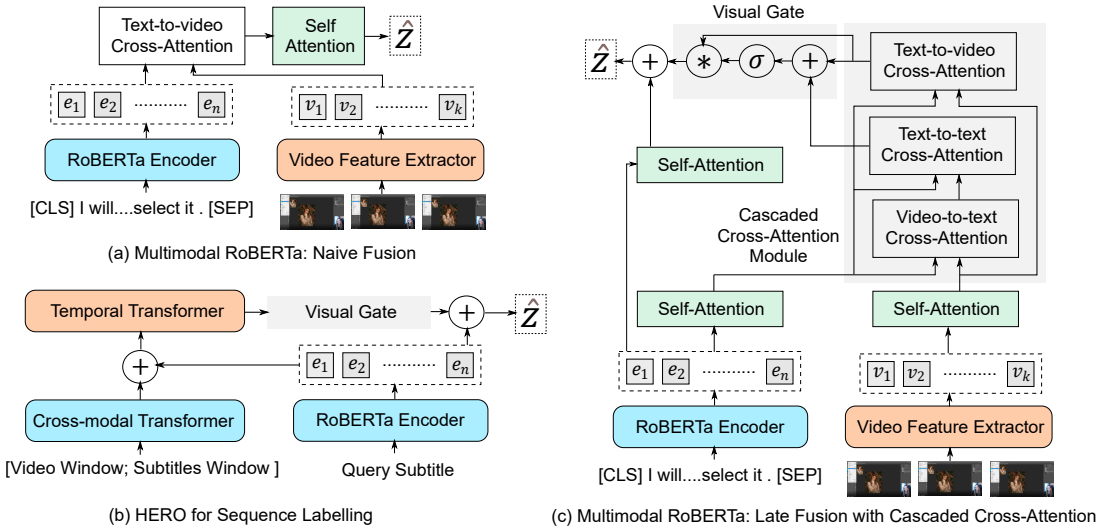


Figure 4: Demonstration of multimodal models: (a) Multimodal RoBERTa with naïve fusion of video and text encodings; (b) Adaptation of HERO for sequence labelling with late fusion, see (c) for visual gate; (c) Multimodal RoBERTa with cascaded cross-attention and late fusion; σ, $+$, $*$ represent sigmoid function, concatenation and matrix multiplication respectively.

5.3 Cascaded Cross-Attention & Late Fusion

Using the intuition that the text signal would provide candidates for the vision model, which is subsequently used for filtering out the cases without intent, we design the cascade cross-attention model as follows: First, we extract the set of contextualized embeddings E from the text encoder f_{enc} and transform it through two self-attention layers to create a two-stream architecture (see Fig. 4). In the first stream, the text encodings are processed through a single-layer of self-attention to produce E_1. In the second stream, the output from self-attention i.e. E_2, is combined with video embeddings through a cascaded cross-attention module. Let $V = [v_1, v_2, ..., v_k]$ be the input sequence of video embeddings. The cascaded module contains three cross-attention layers: video-to-text $f_{v2t}(\cdot)$, text-to-video $f_{t2v}(\cdot)$ and text-to-text cross-attention $f_{t2t}(\cdot)$, with outputs computed as:

$$S_1 = f_{v2t}(W_m V + b_m, E_2)$$
$$S_2 = f_{t2t}(E_2, S_1)$$
$$S_3 = f_{t2i}(E_{s2}, W_m V + b_m)$$

where W_m, b_m are the parameters of a linear layer for transforming video embeddings. Next, the outputs from cross-attention layers are concatenated, linearly mapped and transformed into 0-1 values using a sigmoid, to generate the visual gate (see Fig. 4 (c)). Finally, the output from cross-attention layer is multiplied with this gate, i.e.

$$S_{gate} = sigmoid(W_g[S_2; S_3] + b_g)$$
$$S_{clf} = [S_{gate} * S_3; E_{s1}]$$

The visual gate is dynamically computed using the contextualized video representations and is used to trim the video signal to the relevant bits. This helps in regulating the contribution of the two modalities for the final prediction as per the input. The concatenation represents the late-fusion of text-only embeddings and video-contextualized text embeddings. This merged representation is then sent to the classifier layer for classification i.e. $\hat{Z} = softmax(W_c * S_{clf} + b_c)$.

5.4 Sequence Labelling with Joint Video-Text Pretraining

In order to leverage joint modelling of video and text modalities through large-scale pretraining, we adapt the pretrained HERO (Li et al., 2020) and ClipBERT (Lei et al., 2021) for sequence tagging.

HERO. For each sample in video-based intent detection, we send the video clip and the corresponding subtitle for context as well as query, as input to HERO. V_{cross} represents the cross-contextualized frame embeddings from the Cross-modal Transformer module, which is then concatenated with query embeddings W^q_{emb} before being sent to the Temporal Transformer f_{temp} in HERO

Type	Model	Video Embeddings	Tool Intents			Creative Intents		
			P	R	F	P	R	F
Unimodal	CRF	-	0.43	0.55	0.48	0.17	0.1	0.13
	RoBERTa	-	0.53	0.65	0.58	0.21	0.39	0.27
	GPT2	-	0.41	0.67	0.51	0.12	0.25	0.17
Multimodal (Unimodal Pretraining)	RoBERTa + Naive Fusion	3D ResNext	0.48	0.62	0.54	0.19	0.48	0.27
		2D ResNet	0.54	0.65	0.59	0.23	0.38	0.28
		SlowFast	0.58	0.65	0.61	0.22	0.40	0.29
	RoBERTa + Late Fusion	3D ResNext	0.55	0.64	0.59	0.23	0.41	0.29
		2D ResNet	0.58	0.62	0.60	0.24	0.26	0.25
		SlowFast	0.60	0.66	0.62	0.24	0.41	0.30
Multimodal Pretraining	HERO	2D ResNet + SlowFast	0.57	0.65	0.61	0.23	0.43	0.30
	HERO + Late Fusion	2D ResNet + SlowFast	0.62	0.61	0.62	0.30	0.31	0.30
	ClipBERT	-	0.53	0.66	0.59	0.19	0.35	0.25
	ClipBERT + Late Fusion	-	0.54	0.67	0.60	0.21	0.29	0.27

Table 4: Partial-match based results on the test split of the Behance Intent Discovery dataset.

for global contextualization. Thus, the output is:

$$S_{temp} = f_{temp}([V^{cross}; W^{q}_{emb}])$$
$$S_{out} = S_{temp}[N_v : (N_v + N_t), :]$$

where N_v and N_t are the number of frames and tokens in video and query respectively. The output of f_{temp} is masked to select the representations pertaining to the query only. In the naïve fusion setting, S_{out} is then sent to the classifier layer.

ClipBERT. Similarly, the output S_{out} from the Cross-modal Transformer f_{cross} in ClipBERT is masked and sent to the classifier layer for prediction i.e. $S_{out} = f_{cross}([V; W^{q}_{emb}])[: N_t, :]$.

Late Fusion. We integrate the late fusion approach into HERO and ClipBERT as follows:

$$S_{gate} = sigmoid(W_g * S_{out} + b_g)$$
$$S_{clf} = [S_{gate} * S_{out}; W^{q}_{emb}]$$

where the visual gate is computed as in Sec. 5.3 (see Fig. 4) and S_{clf} is sent to the classifier layer.

6 Experiments

Evaluation. Since the transcribed phrases in Behance Livestreams are the result of an automatic speech recognition (ASR) system, the exact span match metrics might be distorted by ASR errors. Hence, we use a more lenient 75% partial match-based Precision/Recall/F-score metric i.e., if there is more than 75% overlap between the ground truth and predicted span, we consider it as a match.

Video Representations. We experiment with 3D ResNext-101 (Xie et al., 2017) (fps=6), SlowFast (Feichtenhofer et al., 2019) (clip length=2s) and 2D ResNet-152 (He et al., 2016) (clip length=2s) following preprocessing steps in Li et al. (2020).

Models. We use the RoBERTa$_{LARGE}$ (Liu et al., 2019) models for the unimodal experiments, as well as the multimodal experiments that are based on unimodal pretrained models. We use the pretrained HERO (Li et al., 2020) and ClipBERT (Lei et al., 2021) in the remaining experiments; their language encoders are initialized from pretrained RoBERTa$_{BASE}$ and BERT$_{BASE}$ (Devlin et al., 2019) models. Each model is trained end-to-end using fully-supervised training and is subjected to grid-search based hyper-parameter optimization. The best checkpoints are selected based on overall F-Score. See Appendix for bounds.

7 Results

In this section, we discuss results from various models on the Behance Intent Discovery dataset (see Table 4).

The text baselines. Starting with the text-only baselines, we see the best performance from the RoBERTa models, i.e., 58% and 27% partial match F-scores on the tool and creative intents, respectively (rows 2 and 3 in Table 4). Notably, the tool intent predictor is biased with high recall but low precision performance i.e. it retains too many candidates, many of which do not correspond to any intents. These results also demonstrate that large pretrained language models like RoBERTa and GPT2 struggle to comprehend the abstract ideas represented in creative intents.

The Naïve Fusion models. The Naïve Fusion approach with pretrained RoBERTa yields upto 2% improvement over the text-only baselines. In some cases, such as the 3D ResNext representations, this approach degrades the performance, especially in

the harder creative intent set. We attribute this to the difference in informativeness between the text and the video signal, as discussed in Sec. 5.2.

The Late Fusion models. With the Late Fusion approach, we see significant improvements in almost all cases. Compared to the corresponding Naïve Fusion models, Late Fusion models mainly improve precision for tool intents. This result supports our hypothesis that the video signal is most useful as a gate to filter out non-intent candidates from the textual signal. The SlowFast representations prove especially beneficial for creative intents, as seen in row 9 in Table 4. With the use of multimodal pretrained models like HERO and ClipBERT, we observe significant improvements in prediction of tool intents and smaller improvements for creative intents with a simple adaptation of the prediction head for sequence labelling (see Sec. 5.4). HERO uses video representations from pretrained encoders while ClipBERT operates on raw videos; both approaches work well with the software-based video domain yielding upto 3% and 1% improvement on tool intents respectively (rows 10, 12 in Table 4) over the unimodal RoBERTa models. Larger improvements are seen from further augmenting these models with late fusion i.e. 1% improvement on tool intents (rows 11, 12 in Table 4). The late fusion RoBERTa model using SlowFast features (row 9 in Table 4) performs best for creative intents, with 3% improvement over the text-only baseline.

We see similar trends from experiments on the validation set of the Behance Intent Discovery dataset. See results in Appendix.

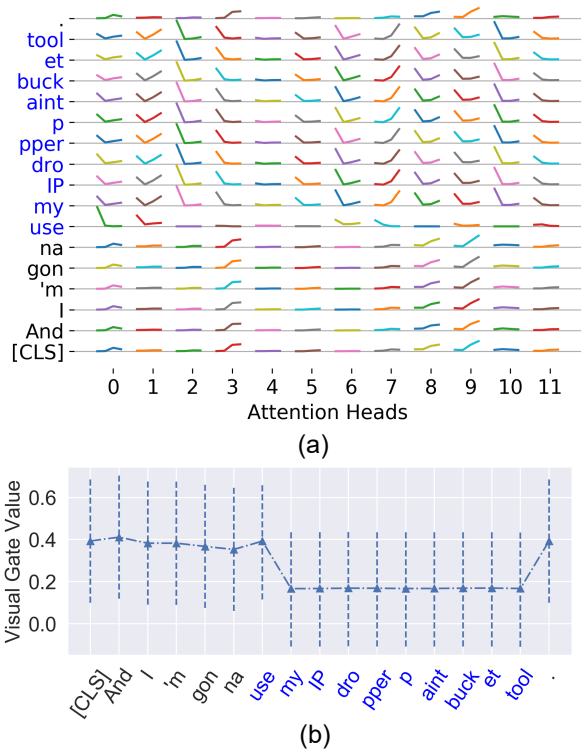


Figure 5: Visualization of (a) temporal attention over video segments from 12 attention heads and, (b) mean ± SD of visual gate values for each token (blue for intent span), using the RoBERTa+Late Fusion model.

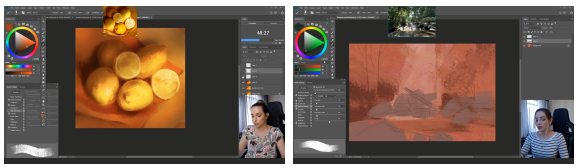


Figure 6: Wrong predictions (red) from unimodal RoBERTa which are solved by adding video signal.

8 Analysis & Discussion

In this section, we perform qualitative analysis of the late fusion approach and examine the effect of video clip length. We also discuss a semi-automated approach to creating annotations for intent extraction and use the data in combination with manual annotations for improved results. See Appendix for more analyses.

8.1 Qualitative Analysis

In order to understand the inner workings of the late fusion architecture, we examine the cross-attention and visual gate modules of the RoBERTa+Late Fusion model trained with 2D ResNet features. Each row of the attention score matrix $M \in R^{n \times f}$ (for n tokens and f video segments) in text-to-video cross-attention module corresponds to the temporal attention over video clips (represented by a sequence of ResNet feature vectors) for a given token. We plot this score matrix for the 12 attention heads in the RoBERTa model in Figures 5(a) and 7(a). The attention heads are activated in the intent region suggesting a strong interaction between two modalities in important segments of the video.

To understand how the video signal helps the prediction, we first plot the mean and standard deviation of visual gate values (S_{gate}) for each token in Figures 5(b) and 7(b). Results show that the visual gate preferentially relies on the video modality for tokens outside the intent span. Furthermore, in Fig. 6, we show example phrases where the text

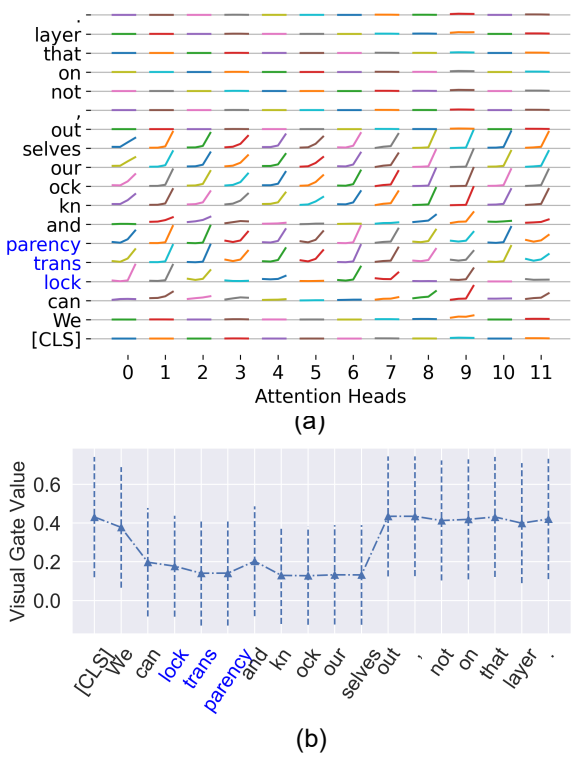


Figure 7: Visualization of (a) temporal attention over video segments from twelve attention heads and (b) mean and standard deviations for the distribution of visual gate values for each token, using the best RoBERTa+Late Fusion model.

only model classifies wrongly as intent while the joint model does not. The phrases themselves appear to be intent but the lack of action in the visual frame indicates that these are chit-chat interactions. Both analyses support our hypothesis that the late fusion model utilizes the video signal to filter intent candidates and improve precision.

8.2 Video Clip Length

As we discuss in Sec. 4, the video clip durations for the tool and creative intents are not specified. We observe that the intended action can span anywhere between 1 second to several minutes. Longer clip lengths are relevant for many creative intents like "make it into something fantasy", "add the arm to this little guy", etc. Hence, we experiment with various clip lengths (10, 20 and 60 secs), but find that larger clip lengths do not lead to further improvements. In fact, with 60 second clips the performance of RoBERTa+Late Fusion model drops below the performance of text-only RoBERTa. This issue could be alleviated with long-range video understanding models (Sener et al., 2020).

8.3 Semi-automated Intent Annotations

Since manual annotation of procedural intents is time-intensive and expensive, we explore a semi-automatic pipeline for creation of intent annotations. The Behance Livestreams corpus contains tool timelines for each livestream, which enumerates the tools used within the software at different points in the livestream. We compute the tf-idf scores for co-occurrence of 896, 287 action-object phrases (from dependency parses of sentences) and corresponding tools in the tool timelines, in order to find the phrases that are frequently used for describing particular tool actions, such as "grab the smudge tool". After filtering the phrases for those with high tf-idf scores, the pool of intent candidates was further cleaned manually, resulting in a final set of 3,697 tool intent candidates. Using this pool of candidates, 24,300 phrases from the Behance Livestreams corpus were identified as tool intent samples. Since it is not straightforward to extract creative intents using similar methods, we first identified key phrases for creative intents from the set of action-object phrases with high term frequency. We then subjected it to manual cleaning (two annotators per sample; κ=0.986) followed by embedding similarity to select creative intents (see Appendix for full pipeline). Using this method, we recovered 7,135 phrases containing creative intents.

We use these semi-automatically collected annotations as additional training data in our experiments with Late Fusion RoBERTa models. Since the manually annotated Behance Intent Discovery dataset is skewed towards negative samples i.e. <25% samples contain intent, we balance the training data by adding 5,000 samples (containing tool or creative intents) from the aforementioned semi-automatically annoated dataset to it. With this balanced data, we see upto 2% improvement in the Late Fusion RoBERTa models. See Appendix.

9 Conclusion

In this paper, we explore the novel task of video-based multimodal intent discovery. We present the unlabelled Behance Livestream corpus consisting of instructional videos for software tools, and the Behance Intent Discovery dataset annotated with tool and creative intents. We propose a late-fusion approach for integration of the video signal with the text signal in a controlled manner for this task, and show significant improvements with unimodal and multimodal pretrained models.

10 Acknowledgements

We would like to thank Tracy King for her detailed feedback and Hailin Jin for making the Behance transcript available. We would also like to thank the reviewers for their useful feedback. This work was partially done while AM was interning at Adobe Research and later extended at UNC, where it was supported by ARO Award W911NF2110220 and DARPA KAIROS Grant FA8750-19-2-1004. The views contained in this article are those of the authors and not of the funding agency.

11 Ethics/Broader Impacts

From an ethics standpoint, we provide a detailed overview of the methods used to create the Behance Livestreams corpus and Behance Intent Discovery dataset in Sec. 4 and more details in the Appendix. We also provide some analyses of the data in Table 3. All of the language data consists of simple English sentences. The dataset comprises livestreamed video tutorials by users of the Behance platform. Behance users grant full usage rights of their content and agree to not hold copyright claims on content in the livestreams videos or transcripts. This content is being made available for free distribution for academic research purposes only and does not allow for redistribution. Aside from the name of the instructor in each video (which is public information), real names of livestream session users or other identifying information does not appear in any of the transcripts. We provide full descriptions of the models used in this paper in Sec. 5. Detailed hyperparameters and bounds for hyperparameter search are included in the Appendix.

Video-based intent discovery serves to enhance the information exploration experience of users on any video-based platform. Since we focus on extracting procedural intent relevant to the goal of the video and in the software domain, we do not anticipate this technology to cause any harm to users, or have any unintended consequences.

References

Swati Agarwal and Ashish Sureka. 2016. Characterizing linguistic attributes for automatic classification of intent based racist/radicalized posts on tumblr micro-blogging website. In *European Intelligence and Security Informatics Conference (EISIC) 2016*.

Alessio Palmero Aprosio, Stefano Menini, and Sara Tonelli. 2020. Creating a multimodal dataset of images and text to study abusive language. *arXiv preprint arXiv:2005.02235*.

Ruichu Cai, Binjun Zhu, Lei Ji, Tianyong Hao, Jun Yan, and Wenyin Liu. 2017. An cnn-lstm attention approach to understanding user query intent from online health communities. In *2017 ieee international conference on data mining workshops (icdmw)*, pages 430–437. IEEE.

Ying-Yu Chen and Shu-Kai Hsieh. 2020. An analysis of multimodal document intent in instagram posts. In *Proceedings of the 32nd Conference on Computational Linguistics and Speech Processing (ROCLING 2020)*, pages 193–207.

Zhiyuan Chen, Bing Liu, Meichun Hsu, Malu Castellanos, and Riddhiman Ghosh. 2013. Identifying intention posts in discussion forums. In *Proceedings of the 2013 conference of the North American chapter of the association for computational linguistics: human language technologies*, pages 1041–1050.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *ICML*.

Alice Coucke, Alaa Saade, Adrien Ball, Théodore Bluche, Alexandre Caulier, David Leroy, Clément Doumouro, Thibault Gisselbrecht, Francesco Caltagirone, Thibaut Lavril, et al. 2018. Snips voice platform: an embedded spoken language understanding system for private-by-design voice interfaces. *arXiv preprint arXiv:1805.10190*.

Deborah A Dahl, Madeleine Bates, Michael K Brown, William M Fisher, Kate Hunicke-Smith, David S Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the atis task: The atis-3 corpus. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

C Ailie Fraser, Joy O Kim, Alison Thornsberry, Scott Klemmer, and Mira Dontcheva. 2019. Sharing the studio: How creative livestreaming can inspire, educate, and engage. In *Proceedings of the 2019 on Creativity and Cognition*, pages 144–155.

Vineet Gupta, Devesh Varshney, Harsh Jhamtani, Deepam Kedia, and Shweta Karwa. 2014. Identifying purchase intent from social posts. In *Eighth International AAAI Conference on Weblogs and Social Media*.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778.

Deng Huang, Peihao Chen, Runhao Zeng, Qing Du, Mingkui Tan, and Chuang Gan. 2020. Location-aware graph convolutional networks for video question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 11021–11028.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Menglin Jia, Zuxuan Wu, Austin Reiter, Claire Cardie, Serge Belongie, and Ser-Nam Lim. 2021. Intentonomy: a dataset and study towards human intent understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12986–12996.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. The hateful memes challenge: Detecting hate speech in multimodal memes. *Advances in Neural Information Processing Systems*, 33.

Junyeong Kim, Minuk Ma, Kyungsu Kim, Sungjin Kim, and Chang D Yoo. 2019. Progressive attention memory network for movie story question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8337–8346.

Kyung-Min Kim, Min-Oh Heo, Seong-Ho Choi, and Byoung-Tak Zhang. 2017. Deepstory: video story qa by deep embedded memory networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 2016–2022.

Christoph Kofler, Martha Larson, and Alan Hanjalic. 2014. Intent-aware video search result optimization. *IEEE transactions on multimedia*, 16(5):1421–1433.

Christoph Kofler, Martha Larson, and Alan Hanjalic. 2016. User intent in multimedia search: a survey of the state of the art and future challenges. *ACM Computing Surveys (CSUR)*, 49(2):1–37.

Julia Kruk, Jonah Lubin, Karan Sikka, Xiao Lin, Dan Jurafsky, and Ajay Divakaran. 2019. Integrating text and image: Determining multimodal document intent in instagram posts. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4622–4632.

Jie Lei, Linjie Li, Luowei Zhou, Zhe Gan, Tamara L Berg, Mohit Bansal, and Jingjing Liu. 2021. Less is more: Clipbert for video-and-language learning via sparse sampling. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7331–7341.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379.

Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. 2020a. What is more likely to happen next? video-and-language future event prediction. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8769–8784.

Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2020b. Tvr: A large-scale dataset for video-subtitle moment retrieval. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XXI 16*, pages 447–463. Springer.

Linjie Li, Yen-Chun Chen, Yu Cheng, Zhe Gan, Licheng Yu, and Jingjing Liu. 2020. Hero: Hierarchical encoder for video+ language omni-representation pre-training. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2046–2065.

Jingzhou Liu, Wenhu Chen, Yu Cheng, Zhe Gan, Licheng Yu, Yiming Yang, and Jingjing Liu. 2020. Violin: A large-scale dataset for video-and-language inference. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10900–10910.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Huaishao Luo, Lei Ji, Botian Shi, Haoyang Huang, Nan Duan, Tianrui Li, Jason Li, Taroon Bharti, and Ming Zhou. 2020. Univl: A unified video and language pre-training model for multimodal understanding and generation. *arXiv preprint arXiv:2002.06353*.

Tegan Maharaj, Nicolas Ballas, Anna Rohrbach, Aaron Courville, and Christopher Pal. 2017. A dataset and exploration of models for understanding video data through fill-in-the-blank question-answering. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6884–6893.

S Meyer. 2015. Linkedin's blockbuster deal with lynda. com: What it means to the online learning industry/s. *Retrieved February*, 15:2016.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

Saiteja Nalla, Mohit Agrawal, Vishal Kaushal, Ganesh Ramakrishnan, and Rishabh Iyer. 2020. Watch hours in minutes: Summarizing videos with user intent. In *European Conference on Computer Vision*, pages 714–730. Springer.

Mingtao Pei, Yunde Jia, and Song-Chun Zhu. 2011. Parsing video events with goal inference and intent prediction. In *2011 International Conference on Computer Vision*, pages 487–494. IEEE.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners.

Fadime Sener, Dipika Singhania, and Angela Yao. 2020. Temporal aggregate representations for long-range video understanding. In *European Conference on Computer Vision*, pages 154–171. Springer.

Aaron Smith, Skye Toor, and Patrick Van Kessel. 2018. Many turn to youtube for children's content, news, how-to lessons. *Pew Research Centre*, 7.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Zineng Tang, Jie Lei, and Mohit Bansal. 2021. Decembert: Learning from noisy instructional videos via dense captions and entropy minimization. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2415–2426.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640.

Hans Van der Meij. 2017. Reviews in instructional video. *Computers & education*, 114:164–174.

Nikhita Vedula, Nedim Lipka, Pranav Maneriker, and Srinivasan Parthasarathy. 2019. Towards open intent discovery for conversational text. *arXiv preprint arXiv:1904.08524*.

Jinpeng Wang, Gao Cong, Xin Wayne Zhao, and Xiaoming Li. 2015. Mining user intents in twitter: A semi-supervised approach to inferring intent categories for tweets. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*.

Di Wu, Liang Ding, Fan Lu, and Jian Xie. 2020. Slotrefine: A fast non-autoregressive model for joint intent detection and slot filling. *EMNLP*.

Congying Xia, Chenwei Zhang, Xiaohui Yan, Yi Chang, and S Yu Philip. 2018. Zero-shot user intent detection via capsule neural networks. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3090–3099.

Saining Xie, Ross Girshick, Piotr Dollár, Zhuowen Tu, and Kaiming He. 2017. Aggregated residual transformations for deep neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1492–1500.

Zekun Yang, Noa Garcia, Chenhui Chu, Mayu Otani, Yuta Nakashima, and Haruo Takemura. 2020. Bert representations for video question answering. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1556–1565.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B Tenenbaum. 2020. Clevrer: Collision events for video representation and reasoning. In *ICLR*.

Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 3342–3352.

Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. 2020. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 12870–12877.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755.

A Dataset

For the semi-automatically created annotations described in Sec. 8.3, we empirically select a window of 10 seconds for computing the scores and retain intent phrases with a term frequency of 5 or higher in the corpus and tf-idf scores of 0.3 or higher with one or more tools. See the full semi-automated pipeline of dataset creation in Fig. 8.

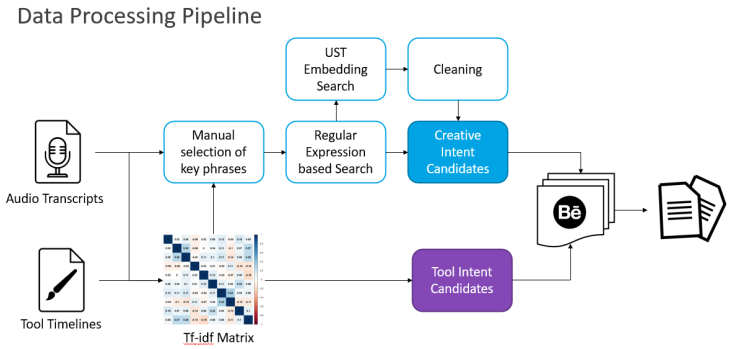


Figure 8: Semi-automated data processing pipeline.

Model	Video Embeddings	Acc.
3-Layer MLP	3D ResNext	0.706
	SlowFast	0.743
3D ResNext-101	-	0.762

Table 5: Results from pilot experiments on usefulness of video modality for multimodal intent discovery.

Probing Experiments. We conducted pilot experiments to probe the usefulness of video signals for intent detection in Behance Livestreams. We prepare a video-only classification dataset for intent classification containing 3,000 samples each for creative, tool and no-intents. We use off-the-shelf ResNext features with a 3-layer MLP classifier as well as finetune ResNext on this task. Using only pretrained video representations, the 3-layer MLP classifier was able to detect the presence of an intent with 70% and 74% accuracy using 3D ResNext and SlowFast features, respectively, while 66% being the chance baseline. With finetuned ResNext, the accuracy improved to 76%. However, the accuracy of classifying between tool and creative intents remained close to random for all models, suggesting the complex nature of creative intents. See Table 5.

B Experiments

For HERO and ClipBERT models, we use the recommended hyperparameters for finetuning in their Github repository.[6,7] For RoBERTa-based models, see the hyperparameters common to all models in Table 8. We performed grid-search based optimization of the variable hyperparameters using the bounds in Table 8. The best performing batch size for all models was found to be 32.

[6] https://github.com/linjieli222/HERO
[7] https://github.com/jayleicn/ClipBERT

C Results

See partial match results for the validation split of Behance Intent Discovery in Table 6.

D Analysis

D.1 Finetuned Video Representations

We see large improvements with sparsely-sampled 2D ResNet video embeddings (see Table 4 which are extracted from ResNet pretrained on the ImageNet dataset. This begs the question, if larger improvements can be had by finetuning the feature extractors on the domain of Behance Livestreams. To facilitate this, we create a dataset of 10,000 images containing snapshots of video livestreams and classified them into one of 50 tool categories using the tool timeline. We finetune ResNet-152 on this dataset with a resulting classification accuracy of 47%. We use the finetuned ResNet to extract sparsely sampled video embeddings and re-run the late fusion experiment with RoBERTa. We see 2% improvement for tool intents and 1% drop in performance on creative intents. This suggests that finetuning feature extractors on the target domain can be beneficial for low-level intents.

D.2 Semi-automated Intent Annotations

As discussed in Sec. 8.3, we use semi-automatically collected annotations as additional training data in our experiments with Late Fusion RoBERTa models. Since the manually annotated Behance Intent Discovery dataset is skewed towards negative samples i.e. <25% samples contain intent, we balance the training data by adding 5,000 samples (containing tool or creative intents) from the aforementioned semi-automatically annoated dataset to it. With this balanced data, we see upto 2% improvement in the Late Fusion RoBERTa models as seen in Table 7. However, with increasing amount of

Type	Model	Video Embeddings	Tool Intents			Creative Intents		
			P	R	F	P	R	F
Unimodal	CRF	-	0.36	0.57	0.44	0.16	0.09	0.12
	RoBERTa	-	0.48	0.78	0.59	0.34	0.52	0.41
	GPT2	-	0.40	0.61	0.48	0.15	0.19	0.15
Multimodal (Unimodal Pretraining)	RoBERTa + Naïve Fusion	3D ResNext	0.46	0.75	0.57	0.3	0.34	0.32
		2D ResNet	0.47	0.77	0.59	0.34	0.64	0.44
		SlowFast	0.52	0.76	0.62	0.36	0.56	0.44
	RoBERTa + Late Fusion	3D ResNext	0.48	0.78	0.59	0.34	0.52	0.41
		2D ResNet	0.48	0.78	0.59	0.34	0.52	0.41
		SlowFast	0.54	0.77	0.62	0.38	0.60	0.44
Multimodal Pretraining	HERO	2D ResNet + SlowFast	0.51	0.72	0.6	0.34	0.31	0.33
	HERO + Late Fusion	2D ResNet + SlowFast	0.56	0.73	0.63	0.37	0.53	0.43
	ClipBERT	-	0.53	0.71	0.61	0.28	0.47	0.35
	ClipBERT	-	0.56	0.73	0.63	0.31	0.48	0.37

Table 6: Partial-match based results on the validation split of the Behance Intent Discovery dataset.

Model	Video Embeddings	Dataset	Tool Intents			Creative Intents		
			P	R	F	P	R	F
RoBERTa + Late Fusion	SlowFast	20K Manual Only	0.60	0.66	0.62	0.24	0.41	0.30
	SlowFast	20K Manual + 5K Semi	0.59	0.69	0.64	0.20	0.44	0.29
	SlowFast	20K Manual + 10K Semi	0.41	0.70	0.51	0.19	0.43	0.25

Table 7: Partial-match based results on the test split of the Behance Intent Discovery dataset using manual annotations and semi-automatically created annotations.

Hyperparameter	Value
Common Hyperparameters	
#Training Epochs	10
Max Gradient Norm	1.0
Weight Decay	0.0
Max. Sequence Length	70
Seed	0
Warmup Steps	200
LR Decay	Linear
Optimizer	AdamW (ϵ=1e-8, β_1=0.9, β_2=0.98)
Search Bounds	
Learning Rate	[1e-4, 1e-5, 5e-5, 1e-6, 5e-6]
Batch Size	[8, 16, 32]

Table 8: Common Training Hyperparameters and Search Bounds for RoBERTa models

semi-automatically data, we drastic decline in the precision of the model for both tool and creative intents (see row 3 in Table 7). With the use of better methods for filtering out the useful signal from the noisy data, there might be better results with semi-automatically created annotations. This line of research is important because it promotes scalable annotations which can cover a diverse population of livestreamers from many livestream videos.