

YouMakeup: A Large-Scale Domain-Specific Multimodal Dataset for Fine-Grained Semantic Comprehension

Weiying Wang Yongcheng Wang Shizhe Chen Qin Jin *

School of Information, Renmin University of China

{wy.wang, 2015202012, cszhe1, qjin}@ruc.edu.cn

Abstract

Multimodal semantic comprehension has attracted increasing research interests in recent years, such as visual question answering and caption generation. However, due to the data limitation, fine-grained semantic comprehension which requires to capture semantic details of multimodal contents has not been well investigated. In this work, we introduce “YouMakeup”, a large-scale multimodal instructional video dataset to support fine-grained semantic comprehension research in specific domain. YouMakeup contains 2,800 videos from YouTube, spanning more than 420 hours in total. Each video is annotated with a sequence of natural language descriptions for instructional steps, grounded in temporal video range and spatial facial areas. The annotated steps in a video involve subtle difference in actions, products and regions, which require fine-grained understanding and reasoning both temporally and spatially. In order to evaluate models’ ability for fine-grained comprehension, we further propose two groups of tasks including generation tasks and visual question answering tasks from different aspects. We also establish a baseline of step caption generation for future comparison. The dataset will be publicly available at <https://github.com/AIM3-RUC/YouMakeup> to support research investigation in fine-grained semantic comprehension.

1 Introduction

Videos which naturally contain rich multimodal semantic information have been one of the main sources for knowledge acquisition. In recent years, video semantic comprehension has attracted much research attention, with a number of datasets and tasks being proposed such as activity recognition (Xu et al., 2017), dense video captioning

(Krishna et al., 2017a), visual question answering (Lei et al., 2018, 2019) etc. However, most works are limited to only capturing coarse semantic information such as action recognition in broad categories. Fine-grained comprehension instead has not been fully explored, especially for discriminating actions with subtle difference or understanding temporal relations of actions in a certain activity.

Instructional videos, which contain series of steps to accomplish certain tasks, are suitable sources to investigate fine-grained semantic comprehension and reasoning. As shown in Table 1, different instructional video datasets have been released. However, current datasets suffer from small data scales or coarse annotations to support fine-grained analysis. For example, datasets collected in (Alayrac et al., 2016; Rohrbach et al., 2012; Stein and McKenna, 2013; Kuehne et al., 2014; Das et al., 2013) only contain hundreds or fewer videos and actions. Although COIN dataset (Tang et al., 2019) is in large scale, it aims to cover wide range of action categories instead of distinguishing actions with subtle difference, and also lacks fine-grained step annotations. The YouCook2 dataset (Zhou et al., 2018) contains fairly large number of videos and temporal grounded sentences in the cooking domain. However, since different cooking steps contain apparent visual variation in actions, food and kitchen utilities, it might not require fine-grained reasoning over temporal and spatial dimensions to identify different steps.

In order to overcome previous limitations, we collect a new instructional video dataset named “YouMakeup” in specific makeup domain for fine-grained multimodal semantic comprehension. The advantages of opting for the makeup domain are threefolds. Firstly, makeup instructional videos are more fine-grained in nature because different steps share the same facial background but con-

* Corresponding author.

Dataset	# of videos	Total len(h)	Avg len(m)	Step Annotation			Domain	Source	
				# of steps	Type	T.G			S.G
“5task” (Alayrac et al., 2016)	150	5	2	-	Sent.	✓	-	General	YouTube
COIN (Tang et al., 2019)	11,827	476.5	2.6	46,354	Ctg.(778)	✓	-	General	YouTube
MPII (Rohrbach et al., 2012)	44	8	10.9	5,609	Ctg.(65)	✓	✓	Cooking	recorded
YouCook (Das et al., 2013)	88	2.5	1.6	-	-	-	-	Cooking	YouTube
50 Salads (Stein and McKenna, 2013)	50	4.5	6.4	966	Ctg.(17)	✓	-	Cooking	recorded
Breakfast (Kuehne et al., 2014)	1,989	77	2.3	8,456	Ctg.(48)	✓	-	Cooking	recorded
Ikea FA (Toyer et al., 2017)	101	4	2.3	1,911	Ctg.(-)	✓	✓	Furniture	recorded
YouCook2 (Zhou et al., 2018)	2,000	176	5.3	3,829	Sent.	✓	-	Cooking	YouTube
EPIC-KITCHENS (Damen et al., 2018)	432	55	7.6	39,596	Sent.	✓	✓	Cooking	recorded
YouMakeup (ours)	2,800	421	9	30,626	Sent.	✓	✓	Makeup	YouTube

Table 1: Comparison of different instructional video datasets. Ctg. = pre-defined categories; Sent. = sentence; T.G = temporal grounding; S.G = spatial grounding; ‘recorded’ means videos are self-recorded by collectors. The YouMakeup dataset is unique in large-scale data size and fine-grained annotations.

tain at least one subtle but critical difference in action, tool or facial area. Therefore, it requires fine-grained discrimination within temporal and spatial context. Secondly, there are abundant makeup instructional videos on the Internet with manual commentary or scripts, which makes it easy to collect and annotate. Last but not least, makeup video analysis is of great value which can facilitate both editing and searching process for cosmetic companies and users. The collected YouMakeup dataset consists of 2,800 makeup videos crawled from YouTube, which spans more than 420 hours. As shown in Figure 1, we manually annotate a sequence of natural language sentences to describe different instructional steps for each video and each step is grounded both in temporal video segment and spatial face areas in fine-grained details. There are totally 30,626 steps with 10.9 steps on average for each video, indicating the complexity of makeup activities.

For the purpose of comprehensively evaluating fine-grained analysis, we propose two groups of potential semantic comprehension tasks on YouMakeup: Generation and Question Answering (QA) tasks. The Generation tasks include temporal step segmentation, step caption generation and spatial area grounding, which reflects an overall semantic comprehension performance. In order to further measure video semantic reasoning ability, we design four QA tasks for detailed evaluations from four aspects as illustrated in Figure 8. The Facial Image Ordering task aims to track subtle changes on facial appearance after each step, which requires to reason influences of actions on object states. The Step Ordering task is to sort step descriptions according to their temporal order in the video, requiring temporal action reasoning and visual semantic matching. The Time Range

Selection task requires the precise temporal localization of specific step in the video, forcing models to distinguish fine-grained difference between makeup steps. The Theme Inference task aims to select a best theme for the video, which demands high-level summarization of video content.

The main contributions of this work are three-folds: 1) We introduce a large-scale fine-grained instructional video dataset “YouMakeup” in makeup domain to support research on fine-grained multimodal semantic comprehension. To the best of our knowledge, it is the largest instructional video dataset in specific domain with fine-grained temporal and spatial grounded annotation. 2) We propose two groups of tasks to evaluate fine-grained video comprehension abilities, including generation tasks and four QA tasks, which require fine-grained semantic understanding and reasoning in different aspects and levels. 3) We propose a baseline framework for the step caption generation task to demonstrate that the fine-grained analysis and long temporal dependencies are essential for multimodal semantic comprehension.

2 Related Work

2.1 Instructional Video Datasets

Existing instructional video datasets can be divided into two groups according to the domain diversity as summarized in Table 1. The first group aims to involve diverse activities from different domains. (Alayrac et al., 2016) contains 5 tasks such as “Making a coffee” and “Changing car tire”. COIN (Tang et al., 2019) is a large scale dataset which contains videos of 180 different tasks in 12 domains related to our daily life. These datasets are constructed to improve model’s generalization ability rather than support the fine-grained semantic comprehension. The

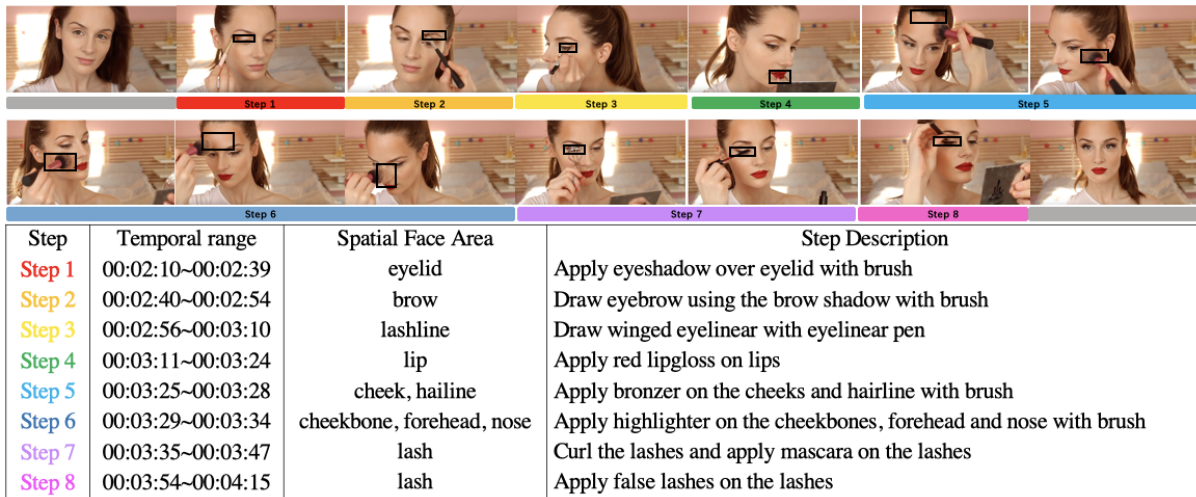


Figure 1: An example video in YouMakeup dataset. We annotate a sequence of step descriptions grounded in temporal video range and spatial face areas for each video. Best viewed in color.

other group focuses on specific domain, such as furniture assembling and cooking. (Rohrbach et al., 2012; Stein and McKenna, 2013; Kuehne et al., 2014) contain videos about simple cooking activities. YouCook (Das et al., 2013) consists of 88 videos with long text summarization. These datasets are limited in both number of videos and actions. YouCook2 (Zhou et al., 2018) is relatively large with 2000 videos, spanning 176 hours. Though cooking events are of rich semantics containing various foods, kitchen utilities and actions (Nishimura et al., 2019; Hahn et al., 2018), such variety makes it hard to measure fine-grained comprehension ability of models. For example, objects in “Sprinkle salt and pepper to the taste” and “Place the bacon at the top” are very different so that it might not be necessary to understand the whole details of actions to distinguish the two steps.

The strengths of our YouMakeup dataset compared with previous works are in two aspects: (1) it is large in scale with 420 hours in total. To the best of our knowledge, it is the largest instructional dataset in specific domain with rich fine-grained annotations. (2) Facial makeup is suitable for fine-grained comprehension in nature for all activities occur on the local facial area with subtle differences.

2.2 Video Comprehension Tasks

A wide range of tasks have been proposed for semantic comprehension on videos, such as action detection, dense video captioning and video question answering etc. The general video captioning

task requires to generate a single sentence for the whole video, which cannot describe video content in details especially for long videos. So in the dense video captioning task, the model needs to detect meaningful events in the video and generate sentence to describe each one. Comparing to the dense captioning tasks which focus on activities with very different actions such as ActivityNet challenge (Krishna et al., 2017a), makeup instructional videos are more fine-grained which contain actions with subtle difference, providing more challenges for semantic comprehension.

Question answering is another way to effectively evaluate semantic understanding. Apart from image based QA datasets such as (Malinowski and Fritz, 2014; Antol et al., 2015; Ren et al., 2015a; Johnson et al., 2017), several video based datasets have been released to explore spatial and temporal inference of the video content. However, they mainly focus on comprehension within short video clips which contain simple activities and interactions, such as (Jang et al., 2017; Kim et al., 2016; Tapaswi et al., 2016; Maharaj et al., 2017), etc. The TVQA dataset (Lei et al., 2018, 2019) is constructed from complex TV shows, but each question is associated with a short clip up to 90 seconds and more focused on joint understanding of visual and speech content. In comparison, our proposed four QA tasks on YouMakeup dataset are used to evaluate video semantic reasoning abilities from different aspects, such as spatial and temporal understanding for long videos, causality reasoning of actions and

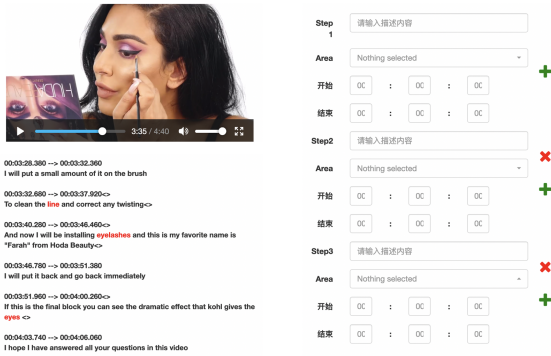


Figure 2: Annotation website

global semantic summarization.

3 YouMakeup Dataset

3.1 Data Collection

Our goal is to build a large-scale multimodal instructional video dataset in the makeup domain to support the fine-grained semantic comprehension research. We start from collecting a list of famous cosmetic brands, such as Chanel, Mac, etc., and beauty bloggers with more than tens of thousands followers. These companies and bloggers are authoritative and professional in the makeup domain with many people learning makeup skills from them. Videos in their official channels are of high quality. Based on the list, we search their official channels on the YouTube and crawl makeup instructional videos together with available meta data such as video id, duration, title, tags and English subtitles generated by YouTube automatically. We process the subtitles into complete sentences aligned with video time stamps.

Since specific names of cosmetic brands are not important to understand makeup procedures, we filter out them in raw titles and subtitles of videos. We first create an initial list of cosmetic brands and then refine it gradually by checking words frequency and finding similar words via word embedding model (Tomas Mikolov, 2013). Finally, we utilize the refined list to remove cosmetic brands in raw texts. We also create lists for facial areas and cosmetic products through the similar process.

3.1.1 Step Annotation

We build an annotation system for step annotation. Figure 2 shows the interface of our annotation system. We provide video and English subtitles in the annotation page to assist the annotation. The products and facial areas in the subtitle are emphasized

in red color to help annotators focus on related information. Annotators are asked to segment the video into a series of steps, which includes labeling the start and end time of each step, selecting the related facial areas in the given facial area list and creating the caption to describe the step according to the video and subtitles. We recruit female college students with more than two years of makeup experience as annotators. Before starting the annotation, each annotator is asked to annotate a test video to verify their capability for annotation. During annotation, each video is annotated by one person and reviewed by another to ensure the annotation quality.

3.1.2 Facial Image Annotation

We extract two groups of facial images from the video. The first group is used to understand the effect of makeup activities on facial appearance, which supports our proposed QA task described in Section 4.2. We extract images at the beginning and the end of each step to capture the facial appearance before and after each step. In order to select images containing faces, we extract images in 40 frames around each time stamp and filter them with a pretrained Multi-Task CNN (MTCNN) (Zhang et al., 2016) for face detection. We then manually filter out unsuitable images such as side face images and pure product images.

The second group is to localize all facial areas referred in each step for spatial grounding. We extract key frames within the annotated segment of each step via comparing similarity of different frames followed by manual filtering of redundant frames. Then we automatically detect facial landmarks in the facial image and align the image region to corresponding facial areas of each step annotated in Section 3.1.1. Finally, we ask annotators to adjust the bounding box of these facial areas on the images and obtain the final grounded facial areas for each step.

3.1.3 Theme Annotation

Inferring the theme of a video is a basic ability to understand the video content. Therefore, we annotate the theme for each makeup video. The original title of the video usually summarizes the content or highlights specific features in the video, which can be treated as the theme of the video. As mentioned above, the specific cosmetic brand names are removed in the title for generalization.



Figure 3: Different categories of the makeup instructional video

We further ask annotators to refine the title with the help of related meta information to make the theme more accurate.

3.2 Dataset Statistics

The final YouMakeup dataset contains 2800 videos, spanning 420 hours 50 minutes, and rich annotations. Videos can be mainly divided into three categories as shown in Figure 3: 1) Makeup for special occasions, such as school days or wedding days; 2) Makeup tips for specific facial area or cosmetic products, such as eye makeup or wearing red lipstick; 3) Makeup transformation, such as celebrities transformation. The first and third types usually create full looks, while the second focuses on specific step or facial area. We split the dataset into training, testing and validation set by 70%: 20%:10% and set up all the tasks on this division.

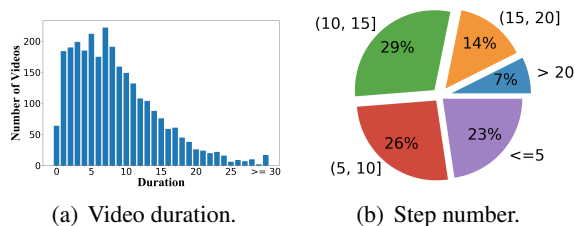


Figure 4: Statistics of video duration and step number.

3.2.1 Video Length

Different from previous datasets containing videos with similar duration (Lei et al., 2018; Tang et al., 2019), videos in YouMakeup dataset are of various lengths, which reflects the complex situation in the realistic world. As shown in Figure 4(a), the length of videos varies from 15s to 1h with

9min on average. The large diversity of length results from the diverse video categories and styles. For example, tutorials from companies are usually short and come straight to the point, while those from beauty bloggers are more complex for they may show makeup skills or share their opinions about products in details.

3.2.2 Makeup Steps

There are 30,626 annotated steps in total with average 10.9 steps per video in YouMakeup. Figure 4(b) shows the distribution of step number. Compared with instructional video datasets in general domains, YouMakeup is more complex with more steps on average and therefore bringing more challenge for semantic understanding.

Each step associates with at least one and up to seven facial areas. The frequency of grounded facial areas are presented in Figure 5. All these areas are close to each other on the face and might contain overlaps for some of them. For example, brow and brow bone are closely adjacent to each other and the under-eye area overlaps with cheeks. Therefore, fine-grained understanding is required to distinguish these areas.

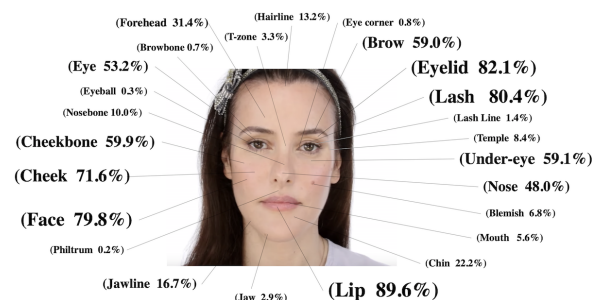


Figure 5: Frequency of grounded facial areas in YouMakeup dataset.

There are more than 1500 unique words occurred in the step captions. Figure 6(a) shows the Wordcloud for most frequent 100 words excluding stopwords. The most frequent words include actions, products, facial areas and tools as summarized in Table 2. These four categories of words can be combined in various ways, generating large number of different fine-grained makeup activities. For example, Figure 7 illustrates the fine-grained activities for the action “apply”. The graph shows that the number of activities is large. Though the activities are similar, they are distinct in actions, products, facial areas or other aspects. Therefore, fine-grained comprehension is required for telling such subtle differences.

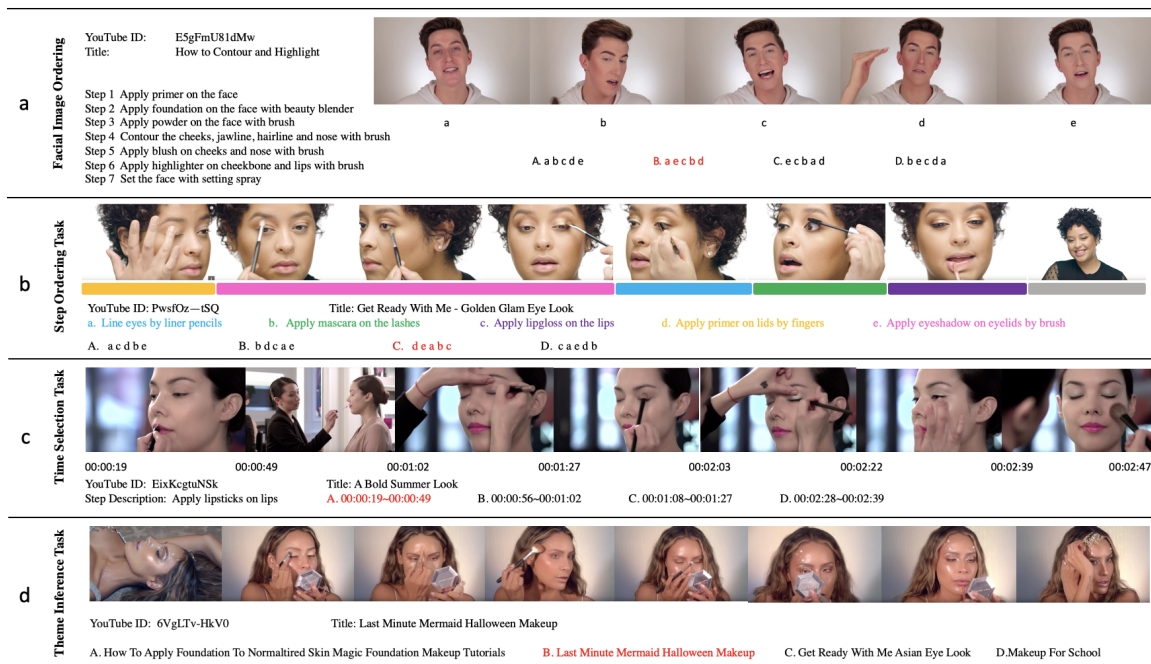


Figure 8: Four VQA tasks on YouMakeup Dataset. Best viewed in color.

answers, including the original one as the ground truth. We finally collect 12,000 questions for this task, including 8,400 for training, 1,200 for validation, and 2,400 for testing.

4.4 Task IV: Time Range Selection

In the time range selection task, models are required to localize a specific step in the video accurately. We set the task in QA form where both the video, the step caption and four candidate answers are provided. As the case shown in Figure 8(c), given the caption, the model needs to select the most accurate time range from four candidate answers after reviewing the whole video. To accomplish this task, models need to jointly understand the step caption and the video, distinguish the fine-grained difference of the makeup activities in each video clips to find the best answer.

To form the question, we first select one step from the video, provide its step caption as question and set its time range as ground truth. Then we choose the other three steps which have some overlap in facial area with the question step from the same video, and set their time range as the candidate answers. We finally collect 12,000 questions for this task, including 8,400 for training, 1,200 for validation and 2,400 for testing.

4.5 Task V: Theme Inference

The theme inference task requires to infer the theme of a video. People can compress complicated content and extract key information. Several tasks have been set up to help the model develop such ability, such as the video summarization task which summarizes the video in several sentences. Theme inference is another way to evaluate this kind of ability. Since a video can be summarized with focus on multiple aspects, it is difficult to treat theme inference as generation task. We set it in QA form based on the annotation in Section 3.1.3.

As shown in Figure 8(d), we provide four candidate answers including the ground-truth. To generate candidate answers, we set up the candidate answer set with titles and tags of all videos. Then we utilize the FastText (Bojanowski et al., 2016), Doc2Vec (Le and Mikolov, 2014) and Word2Vec (Tomas Mikolov, 2013) together for theme feature representation to search the top 10 nearest themes for each video. These themes are then selected by the annotator to form the candidate answers. Given the makeup videos, the model needs to select the most suitable theme from four choices.

Different from the video classification task which divides videos into fixed pre-defined categories, the candidate answers in the theme inference task are natural language sentences different from each other. As shown in Figure 6(b), the

themes involve diverse aspects, such as “smokey” and “natural” for makeup style, “night” and “halloween” for occasions, and “pink” and “red” for color tone. Thus, the theme inference task requires a fine-grained comprehensive understanding of the video content in various aspects.

5 Experiment

We build a step caption generation system to provide a baseline for our YouMakeup dataset and demonstrate the necessity of fine-grained spatial and temporal understanding to solve the task. The system utilizes the groundtruth step segmentation in order to evaluate the captioning ability alone, and is evaluated by the standard captioning metrics including BLEU (Papineni et al., 2002), METEOR (Denkowski and Lavie, 2014), ROUGE (Flick, 2004), CIDEr (Vedantam et al., 2015) and SPICE (Anderson et al., 2016).

Implementation Details Our step caption generation model is based on the encoder-decoder captioning framework, which is widely used in video caption generation (Vinyals et al., 2015; Chen et al., 2019). The encoder converts the video clip into a fixed-dimensional vector and then the decoder generates word sequences conditioning on the encoded vector. Since both spatial and temporal reasoning are important for the task, we propose two types of encoder as follows: 1) spatial encoder: faster RCNN (Ren et al., 2015b) pretrained on the VisualGenome dataset (Krishna et al., 2017b) is used to extract object features in a single frame. We select one frame for each video clip and detect at most 36 objects in the frame. Mean pooling is applied on the extracted object features to generate the video-level representation. 2) temporal encoder: Resnet152 (He et al., 2016) pretrained on the ImageNet dataset (Deng et al., 2009) is used to extract features for each frame. We extract global frame-level features for every 16 frames and apply mean pooling on the temporal dimension to generate the global video-level representation. We employ the LSTM as our decoder, which contains 1 hidden layer with 512 hidden units. Adam optimizer is used to train our model with batch size of 128 and learning rate of 0.0001. We train at most 100 epochs and select the best model according to captioning performance on the validation set.

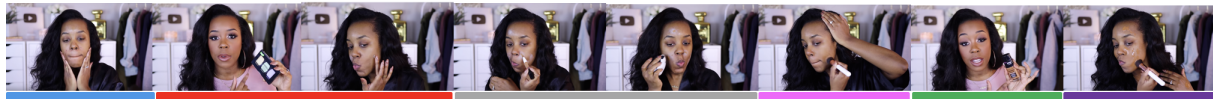
Results. Table 3 presents the step captioning performance with ground truth step segmentation

Table 3: Step captioning performance with groundtruth step segmentation on the YouMakeup dataset.

	B@4	Meteor	Rouge	Cider	Spice
spatial	13.1	19.7	46.3	94.6	27.1
temporal	16.9	21.9	48.8	130.8	34.3
spatial+temporal	17.9	22.4	50.3	134.7	34.5

on the testing set of YouMakeup. We can see that the combination of spatial and temporal features achieves the best performance, demonstrating spatial and temporal information is complementary to generate step captions. The captioning model based on the spatial feature alone is inferior to that based on the temporal feature because our framework does not utilize all clip information. The overall performance on Cider of the baseline step caption generation model is relatively higher than baselines on other video captioning datasets (Krishna et al., 2017a) due to the fine-grained characteristics of the proposed dataset. In the YouMakeup dataset, the styles of step captions are similar which can make it easy for the caption generation system to achieve high scores since evaluation metrics are not aware of the aspect importance in the caption such as detailed tools, actions etc.

Although the generation system achieves high evaluation scores, we find it fails to capture the fine-grained details in the makeup instructional videos. Figure 9 shows the caption generation results of the baseline system using both temporal and spatial features on a specific video for the first 6 steps. According to the captions of step 1 and 3, the system traces out the procedure roughly, lacking of details such as makeup tools and related facial area. The other three step captions indicate the system’s weak ability on fine-grained video comprehension and confirm the fine-grained characteristics of YouMakeup. In step 2, the system mistakes color correction palette for eyeshadow due to their similar appearance. However, the eyeshadow is applied around eyes while color correction palette is used on the face for color correcting. From step 4 to 6, system shows confusion about the procedure of applying foundation and concealer because they are similar in both appearance and usage. System needs to associate products with their usage methods and facial areas they are applied in order to grasp the subtle difference between different makeup activities for generating accurate step description.



Step	Generated Caption	Ground Truth
Step 1	apply primer on face	apply the primer on face with fingers
Step 2	apply eyeshadow on eyelid with brush	apply the color correction palette on face with fingers
Step 3	apply concealer on face with sponge	apply the concealer with sponge on one side of the face
Step 4	apply foundation on face with sponge	blend the concealer with brush on another side of the face
Step 5	apply foundation on face	apply the foundation on skin with brush
Step 6	apply concealer on face with sponge	apply the foundation on face with beauty sponge

Figure 9: Comparison of generated captions and ground-truth captions for different steps. Best viewed in color.

6 Conclusion

In this paper, we introduce a new large-scale instructional video dataset named YouMakeup for fine-grained semantic comprehension. The YouMakeup dataset contains 2,800 makeup instructional videos spanning more than 420 hours in total. Based on the characteristics of makeup instructional videos and the rich annotations of temporal boundaries, grounded facial areas and natural language descriptions of steps, our collected dataset is more suitable to support the fine-grained video comprehension research than previous datasets. We further design a generation task and four question answering tasks to thoroughly evaluate the fine-grained semantic comprehension ability from different aspects and levels. A baseline system for step caption generation also demonstrates the necessity of fine-grained spatial and temporal information. In the future work, we plan to make thorough exploration on these proposed tasks. We will make the dataset publicly accessible in order to support the research investigation in fine-grained semantic comprehension.

Acknowledgments

This work was supported by National Natural Science Foundation of China (No. 61772535), Beijing Natural Science Foundation (No. 4192028), and National Key Research and Development Plan (No. 2016YFB1001202). We would like to thank our group member Jingjun Liang for his help in building the annotation website and all the annotators for their careful annotations.

References

Jean-Baptiste Alayrac, Piotr Bojanowski, Nishant Agrawal, Josef Sivic, Ivan Laptev, and Simon Lacoste-Julien. 2016. Unsupervised learning from narrated instruction videos. In *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition, pages 4575–4583.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. *Adaptive Behavior*, 11(4):382–398.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Shizhe Chen, Yuqing Song, Yida Zhao, Qin Jin, Zhaoyang Zeng, Bei Liu, Jianlong Fu, and Alexander Hauptmann. 2019. Activitynet 2019 task 3: Exploring contexts for dense captioning events in videos. *arXiv preprint arXiv:1907.05092*.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, et al. 2018. Scaling egocentric vision: The epic-kitchens dataset. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 720–736.

Pradipto Das, Chenliang Xu, Richard F Doell, and Jason J Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2634–2641.

Jia Deng, Wei Dong, Richard Socher, Li Jia Li, Kai Li, and Fei Fei Li. 2009. Imagenet: A large-scale hierarchical image database. In *IEEE Conference on Computer Vision Pattern Recognition*.

Michael Denkowski and Alon Lavie. 2014. [Meteor universal: Language specific translation evaluation for any target language](#). In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380. Association for Computational Linguistics.

- Carlos Flick. 2004. Rouge: A package for automatic evaluation of summaries. In *Workshop on Text Summarization Branches Out*.
- Meera Hahn, Nataniel Ruiz, Jean-Baptiste Alayrac, Ivan Laptev, and James M Rehg. 2018. Learning to localize and align fine-grained actions to sparse instructions. *arXiv preprint arXiv:1809.08381*.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Sun Jian. 2016. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*.
- 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.
- Justin Johnson, Bharath Hariharan, Laurens van der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2901–2910.
- K Kim, C Nan, MO Heo, SH Choi, and BT Zhang. 2016. Pororoqa: Cartoon video series dataset for story understanding. In *Proceedings of NIPS 2016 Workshop on Large Scale Computer Vision System*, volume 15.
- Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. 2017a. Dense-captioning events in videos. In *The IEEE International Conference on Computer Vision (ICCV)*.
- Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Li Jia Li, Li Jia Li, and David A. Shamma. 2017b. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision*, 123(1):32–73.
- Hilde Kuehne, Ali Arslan, and Thomas Serre. 2014. The language of actions: Recovering the syntax and semantics of goal-directed human activities. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.
- Quoc Le and Tomas Mikolov. 2014. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196.
- Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. 2018. Tvqa: Localized, compositional video question answering. In *EMNLP*.
- Jie Lei, Licheng Yu, Tamara L Berg, and Mohit Bansal. 2019. Tvqa+: Spatio-temporal grounding for video question answering. *arXiv preprint arXiv:1904.11574*.
- 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.
- Mateusz Malinowski and Mario Fritz. 2014. A multi-world approach to question answering about real-world scenes based on uncertain input. In *Advances in neural information processing systems*, pages 1682–1690.
- Taichi Nishimura, Atsushi Hashimoto, Yoko Yamakata, and Shinsuke Mori. 2019. Frame selection for producing recipe with pictures from an execution video of a recipe. In *Proceedings of the 11th Workshop on Multimedia for Cooking and Eating Activities*, pages 9–16. ACM.
- Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*.
- Mengye Ren, Ryan Kiros, and Richard Zemel. 2015a. Exploring models and data for image question answering. In *Advances in neural information processing systems*, pages 2953–2961.
- Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. 2015b. Faster r-cnn: towards real-time object detection with region proposal networks.
- Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele. 2012. A database for fine grained activity detection of cooking activities. In *2012 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1194–1201. IEEE.
- Sebastian Stein and Stephen J McKenna. 2013. Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *Proceedings of the 2013 ACM international joint conference on Pervasive and ubiquitous computing*, pages 729–738. ACM.
- Yansong Tang, Dajun Ding, Yongming Rao, Yu Zheng, Danyang Zhang, Lili Zhao, Jiwen Lu, and Jie Zhou. 2019. Coin: A large-scale dataset for comprehensive instructional video analysis. *arXiv preprint arXiv:1903.02874*.
- 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.
- Kai Chen Greg Corrado Jeffrey Dean Tomas Mikolov, Ilya Sutskever. 2013. Distributed representations of words and phrases and their compositionality. *arXiv preprint arXiv:1310.4546*.

- Sam Toyer, Anoop Cherian, Tengda Han, and Stephen Gould. 2017. Human pose forecasting via deep markov models. In *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8. IEEE.
- Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*.
- Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *IEEE Conference on Computer Vision Pattern Recognition*.
- Huijuan Xu, Abir Das, and Kate Saenko. 2017. R-c3d: Region convolutional 3d network for temporal activity detection. In *Proceedings of the IEEE international conference on computer vision*, pages 5783–5792.
- K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE Signal Processing Letters*, 23(10):1499–1503.
- Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598.