Comprehensive Information Integration Modeling Framework for Video Titling

Shengyu Zhang[1*], Ziqi Tan[1*], Jin Yu[2], Zhou Zhao[1†], Kun Kuang[1], Tan Jiang[1], Jingren Zhou[2],
Hongxia Yang[2†], Fei Wu[1†]
[1] College of Computer Science and Technology, Zhejiang University
[2] Alibaba Group
{sy_zhang,tanziqi,zhaozhou,kunkuang,jiangtan,wufei}@zju.edu.cn
{kola.yu,jingren.zhou,yang.yhx}@alibaba-inc.com

ABSTRACT

In e-commerce, consumer-generated videos, which in general deliver consumers' individual preferences for the different aspects of certain products, are massive in volume. To recommend these videos to potential consumers more effectively, diverse and catchy video titles are critical. However, consumer-generated videos seldom accompany appropriate titles. To bridge this gap, we integrate comprehensive sources of information, including the content of consumer-generated videos, the narrative comment sentences supplied by consumers, and the product attributes, in an end-to-end modeling framework. Although automatic *video titling* is very useful and demanding, it is much less addressed than video captioning. The latter focuses on generating sentences that describe videos as a whole while our task requires the product-aware multi-grained video analysis. To tackle this issue, the proposed method consists of two processes, *i.e.*, granular-level interaction modeling and abstraction-level story-line summarization. Specifically, the granular-level interaction modeling first utilizes temporal-spatial landmark cues, descriptive words, and abstractive attributes to builds three individual graphs and recognizes the intra-actions in each graph through Graph Neural Networks (GNN). Then the global-local aggregation module is proposed to model inter-actions across graphs and aggregate heterogeneous graphs into a holistic graph representation. The abstraction-level story-line summarization further considers both frame-level video features and the holistic graph to utilize the interactions between products and backgrounds, and generate the story-line topic of the video. We collect a large-scale dataset accordingly from real-world data in Taobao, a world-leading e-commerce platform, and will make the desensitized version publicly available to nourish further development of the research community[1]. Relatively extensive experiments on various datasets demonstrate the efficacy of the proposed method.

CCS CONCEPTS

• **Computing methodologies** → *Hierarchical representations*; **Natural language generation**.

KEYWORDS

Video title generation; Graph neural network; Video recommendation; Mobile E-Commerce

ACM Reference Format:
Shengyu Zhang, Ziqi Tan, Jin Yu, Zhou Zhao, Kun Kuang, Tan Jiang, Jingren Zhou, Hongxia Yang, Fei Wu. 2020. Comprehensive Information Integration Modeling Framework for Video Titling. In *Proceedings of the 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD '20), August 23–27, 2020, Virtual Event, CA, USA.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3394486.3403325

1 INTRODUCTION

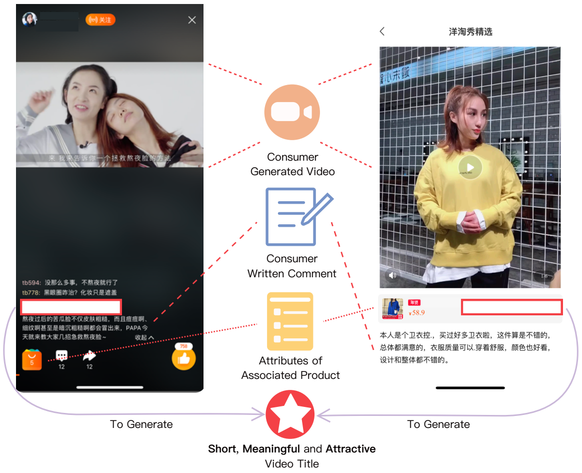
Nowadays, a massive number of short videos are being generated, and short video applications play a pivotal role in the acquisition of new customers [6]. In e-commerce, a great many consumers upload videos in the product comment area to share their unique shopping experiences. Different from professional agency-generated videos (*e.g.*, Ads), which may deliver a wealth of official information, consumer-generated videos depict consumers' individual preference for the different aspects of certain products, which probably tempt others to buy the same products. Therefore, we are more interested in recommending consumer-generated videos to potential consumers. For video recommendation, it is in general appropriate to generate a descriptive and memorable title for the given video to have potential consumers know the benefit they will receive before diving into the full video contents.

Although the consumer-generated videos often accompany the user-written descriptive comments, these comments are infeasible to be directly taken as titles since many commentary sentences often emphasize the consumers' intuitive perception of some shopping experience (e.g., the logistics and the services) irrelevant to the products themselves (e.g., the functionality of products). To this end, we seek an appropriate way to generate titles of consumer-generated videos from the perspective of distilling overall meaningful information depending on the following three kinds of facts: 1) the consumer-generated videos, which visually illustrate the

[1] Dataset available at https://github.com/LightGal/VideoTitling

[*] These authors contributed equally to this work.
[†] Corresponding Authors.
[*] Work was performed when S. Zhang and Z. Tan were interns at Alibaba Group.

KDD '20, August 23–27, 2020, Virtual Event, CA, USA
© 2020 Association for Computing Machinery.
ACM ISBN 978-1-4503-7998-4/20/08…$15.00
https://doi.org/10.1145/3394486.3403325


关注
洋淘秀精选
来 我来告诉你一个拯救熬夜脸的方法
Consumer Generated Video
Consumer Written Comment
Attributes of Associated Product
¥58.9
收起
5
12
12
758
To Generate
To Generate
Short, Meaningful and Attractive Video Title

tb594: 没那么多事，不熬夜就行了
tb778: 黑眼圈咋治? 化妆只是遮盖

熬夜过后的苦瓜脸不仅皮肤粗糙，而且痘痘啊、细纹啊甚至是暗沉粗糙啊都会冒出来，PAPA今天就来教大家几招急救熬夜脸~

本人是个卫衣控.，买过好多卫衣啦，这件算是不错的，总体都满意的，衣服质量可以.穿着舒服，颜色也好看，设计和整体都不错的。

Figure 1: An illustration of task definition and two real application scenarios in Mobile Taobao, *i.e.*, *Trending Video Topics* (left) and *Selected Buyers Show* (right). We aim to generate titles for consumer-generated videos considering the narrative comments and the attributes of associated products.

detailed characteristics and story-line topics of products; 2) the comment sentences written by consumers, which mainly narrate the consumers' preference to different aspects of products as well as consumers' shopping experience; 3) the attributes of associated products, which specify the human-nameable qualities of the products (e.g., functionality, appearance). It can be observed that each of them contains unique and beneficial information that should not be overlooked. We name our task as product-aware *video titling* (the generation of video titles). Figure 1 illustrates two real application scenarios of *video titling* – Trending Video Topics and Selected Buyers Show in Mobile Taobao, which is the largest online e-commerce platform in China.

To the best of our knowledge, we are the initiative to investigate such a problem in a real complex scenario. One related and general task in the literature can be video captioning. The challenging and practical nature of the task lends itself to various models [24, 36, 38, 42]. However, these RNN-Encoder based methods have limitations. On the one hand, these methods model the video solely in the frame-level, which may be suitable for describing videos as a whole (*i.e.*, recognizing main objects and general actions such as " a man is playing basketball") but may fail to recognize the distinguishing features of products as well as the dynamic change of fine-grained landmarks, *i.e.*, different key-points of products. On the other hand, the original encoder-decoder framework of RNNs is incapable of systematically employing and aggregating the three kinds of information: the visual spatial-temporal dynamics in the video, the narrative description in commentary sentences and the human-nameable aspects of products.

To mitigate these problems, we propose a new learning schema named ***G**raph based **V**ideo **T**itl**e** generator* abbreviated as Gavotte. Specifically, Gavotte is hierarchically comprised of two sub-processes, *i.e.*, *granular-level landmark modeling* and *abstraction-level story-line summarization*. The granular-level sub-process represents the three kinds of information (*i.e.*, the consumer-generated video, the narrative commentary sentence, as well as the attributes of the associated product) as three individual graphs, in order to better capitalize on the intra-actions of granular-level cues inside three kinds of information. Furthermore, the sheer difficulty of capturing the inter-actions of different heterogeneous graphs and aggregating these graphs has not been well investigated yet in the literature. To this end, we propose a *Global-local Aggregation* module enhanced with the soft attention mechanism to effectively model the inter-actions and aggregate heterogeneous graphs into a holistic graph representation. Since consumers usually customize their favorite products with the social surrounding and physical environment (e.g., lighting and decorations) in demonstration videos, it is beneficial to appropriately utilize the interactions between products and backgrounds in order to generate better titles. To capture such information in the frame-level and story topic at the video-level, the *abstraction-level story-line summarization* process models the temporal structure of frames using RNNs with the guidance of granular-level features, driven by the empirical observation that the high-level semantics such as the topic and style are closely related to the local details.

To accommodate our research and applied system development, we collect a large-scale video title generation dataset, named *Taobao video titling dataset* (**T-VTD**) from real-world e-commerce scenario. T-VTD contains about 90,000 <video, comment, attributes, title> quadruples, and is several orders of magnitude greater than most general video captioning datasets considering the quantity and total length of videos. For natural language data, T-VTD has an extensive vocabulary and less repetitive information. Notably, compared to captions from existing video datasets, which mostly describe the objects directly, the titles in our datasets depict different levels of information, including the fine-grained characteristics, the main category, and the overall style of the products as well as video topics. These new features of T-VTD pose many new challenges for general video captioning research and other diversified research topics. We make the desensitized version of T-VTD publicly available to promote further investigation and make our model reproducible.

In summary, the main contributions of this work are three-fold.

- We devise a new learning schema for general video captioning by employing the flexible GNNs and the hierarchical video modeling processes, *i.e.*, the *granular-level landmark modeling*, and the *abstraction-level story-line summarization*.
- We advocate investigating the real-world problem in e-commerce, named product-aware *video titling*, and propose the *Global-local Aggregation* module to capture the inter-actions across heterogeneous graphs and aggregate them into a holistic representation.
- We conduct relatively extensive experiments across various measurements, including human evaluation and online A/B test, on a large industry dataset from the Mobile Taobao. We will also release the Taobao dataset to further nourish the research community.

2 RELATED WORKS

2.1 Video Captioning & Title Generation

For traditional video captioning, template-based methods have made substantial improvements for their ease-of-use and reliable performance [13, 17, 31]. Nevertheless, these kinds of methods highly depend on manually and carefully predefined templates. They are also limited in expression ability since only partial output (word roles, such as subject, verb, and object) is generated.

Notably, deep learning based video captioning methods obtained great success and can mitigate the problems above. Most of them rely on frame features pre-extracted by other off-the-shelf general video understanding models [10, 14]. Venugopalan *et al.*[37] proposes to represent the video as mean-pooled frame features. More advanced methods employ the effective sequence-to-sequence encoder-decoder architecture [36], hierarchical RNN design [24], and soft-attention mechanisms [42]. More recently, Wang *et al.*[38] proposes a reconstruction backward-flow to re-generate the input frame features both locally and globally, ensuring that the decoder hidden vectors contain the necessary video information. Livebot [23] employs a Unified Transformer Model to generate live video comments based on frames and surrounding comments. The main differences between our proposed method and the above are mainly two-fold: 1) They solely model the video in the frame-level to recognize the main object and general event for captioning. Gavotte represents landmark-level features among video frames as a video landmark graph to better capture the granular cues of products for product-aware video titling. 2) These models are incapable of modeling the comprehensive sources of facts. We propose the *global-local aggregation* module to model the fine-grained inter-actions between these facts and aggregate them into a holistic representation.

To the best of our knowledge, Zeng *et al.*[44] is currently the only work that investigates video title generation, which aims to capture the most salient events in videos by proposing a highlight detector. However, this design requires highlight moments annotation for training the highlight detector, which does not apply to our task. It may not recognize the fine-grained characteristics due to the frame-level representation, and cannot model the comprehensive sources of input facts, either.

2.2 Video Graph Representation

To exploit special relationships besides the over-explored sequential frame dependencies, there are several lines of works representing video as graphs. Wang *et al.*[40] firstly represents the objects within and across frames as a spatial-temporal graph for video action recognition. Following this, VRD-GCN [27] builds a video graph using a similar setting and design the ST-GCN module to perform information propagation. AGNN [39] views frames as nodes and models the fully-connected relationships using Graph Neural Networks. To capture the video segment level interactions, Zeng *et al.*[45] represents the pre-extracted segment proposals as the graph for action localization. To the best of our knowledge, our work is an early attempt to apply video graph representation for describing videos in natural language. We differ from the above methods by representing the granular-level landmark cues as a graph instead of the object- or frame-level information. Notably, besides the visual graph, we also represent information of other modality and other structures (*i.e.*, the narrative description, and human-namable attributes) as graphs. We design the *global-local aggregation* module to align and aggregate heterogeneous graphs into a holistic representation rather than single graph modeling.

Notation	Description
e, v, G	the edge, vertex and graph
$\mathbf{e}, \mathbf{v}$	the edge weight / the vertex feature
$\mathbf{V}$	the set of vertex features
$\mathbf{p}, \mathbf{r}$	position/type embedding
N	the number of nodes / the length of sequence/set
D	the dimension of features
$\mathbf{L}_i, \mathbf{l}_{i,j}$	landmarks feature set of the ith frame / the jth feature
$\mathbf{X}, \mathbf{x}_i$	the sequence of frame features / the ith frame feature
$\mathbf{W}, \mathbf{b}$	linear transformation / bias term
$\mathbf{H}, \mathbf{h}_t$	the sequence of RNN hidden features / the feature at step t

Table 1: Notations

3 METHODS

3.1 Overview

Given a consumer-generated video comprised of N_v frames, *i.e.*, $\{x_i\}_{i=1,\ldots,N_v}$, a consumer-written comment of N_w words, *i.e.*, $\{w_t\}_{t=1,\ldots,N_w}$ and the N_a-sized attributes set, $\{a_k\}_{k=1,\ldots,N_a}$, of the associated product, the goal is to generate an abstractive and descriptive video title $\{c_m\}_{m=1,\ldots,N_c}$ for e-commerce mobile display. The main challenges of video titling come from the multi-modal and multi-source nature of the input information, and the sheer difficulty of capturing multi-grained cues within the video. Our proposed model recasts the generation procedure as two heuristic sub-processes, namely *granular-level interaction modeling* and *abstraction-level story-line summarization* to hierarchically and comprehensively model the three kinds of information.

In the *granular-level interaction modeling* process, we firstly represent the three kinds of information as individual graphs, including the video landmark graph G_v, the comment graph G_c, and the attributes graph G_a, separately. These graphs focus on particularly fine-grained information. We obtain the intra-actions of the granular-level cues in each graph using the flexible and effective Graph Neural Networks. To exploit the inter-actions and dependencies across graphs, we propose the *global-local aggregation* module, which transforms heterogeneous graphs into an aggregated graph.

In the *abstraction-level story-line summarization* process, the goal is to model the sequential structure of the video in the frame-level. This process is designed to better recognize the main topic and product-background interaction, which are essential to generate grounded and descriptive video titles. The RNN is employed here since it is capable of learning the temporal dependencies of videos

3.2 Graph Representations

3.2.1 *Video Landmark Graph - G_v.*

Similar to [21], we use landmarks to denote the salient parts of products. To effectively exploit the landmarks in each frame and across frames, we represent these landmarks of a video as a spatial-temporal graph. Previous works [27, 39, 40, 45] have demonstrated the superior performance of video-graph modeling in capturing long-range temporal dependencies and the high-order relationships among the frames/objects, and we transfer this idea to the landmark-level. Specifically, we extract landmark features $\mathbf{L}_i = \{\mathbf{l}_{i,j} \in \mathbb{R}^{D_L}\}_{j=1,\ldots,N_l}$ from the ith frame. N_l denotes how many landmarks are detected within a frame. The extraction details can be found in the appendix A.3. Each landmark feature $\mathbf{l}_{i,j}$ is represented as a vector of dimension D_L. Therefore,

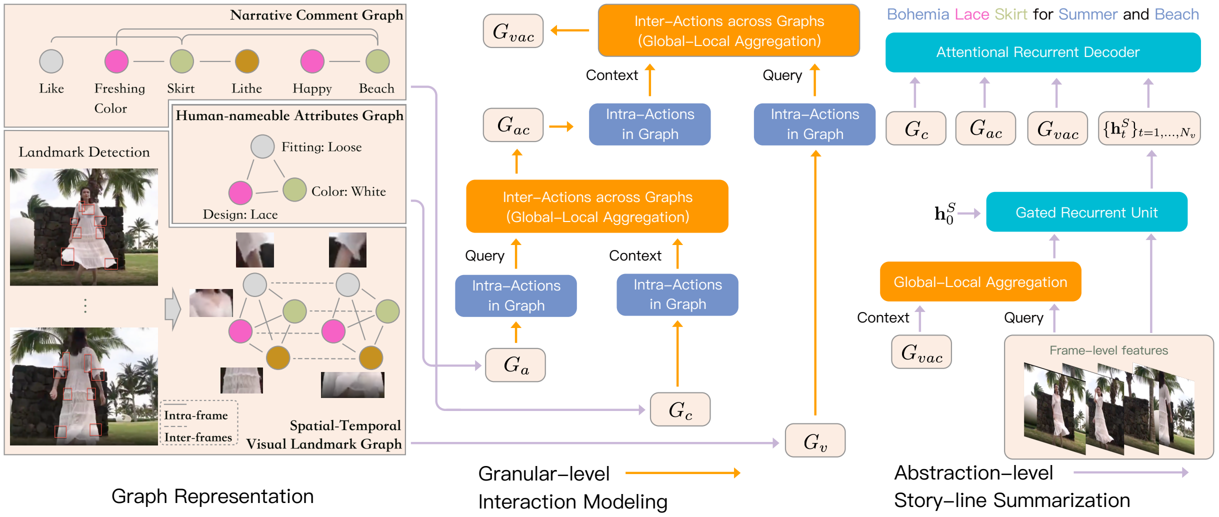


Figure 2: Schematic of our proposed Gavotte, comprising mainly the granular-level interaction modeling process and abstraction-level story-line summarization. The former process focuses more on recognizing the granular-level cues within different facts using graph modeling and combining heterogeneous graphs into a holistic representation, *i.e.*, G_{vac}. As a necessary counterpart, the latter process uses both frame features and the aggregated graph features and is designed to figure out the product-background interaction as well as the story topic. Modules of the same kind are depicted using the same color.

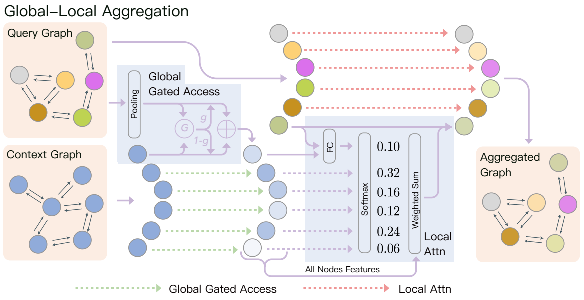


Figure 3: A detailed illustration of the proposed global-local aggregation module for aligning and combining heterogeneous graphs. We firstly encourage context information relevant to the global representation of the query graph (global gated access) and further attend to the context graph for each query node (local attention).

the video landmark graph contains $N_l \times N_v$ nodes (landmarks). We build full connections for two kinds of landmarks: 1) landmarks within the same frame, which help capture the appearance and design of the product as a whole. 2) landmarks of the same kind (*e.g.*, collar) across frames, which help capture the dynamic change and the comprehensive characteristics of each kind of landmark.

However, there is no notion of frame order in this schema (unlike common RNN architectures) and thus losing the temporal dependencies in the input/output. Inspired by the position embedding technique widely used in sequence learning [12, 22], we add a similar position-dependent signal to each landmark feature, which helps capture comprehensive landmark appearance and functionality along the timeline. We further facilitate the landmark feature by adding landmark-type dependent (such as collar and sleeve) embedding. Therefore, the node feature in G_v is initialized as:

$$\mathbf{v}_{i,j}^{l} = \mathbf{l}_{i,j} + \mathbf{p}_i + \mathbf{r}_j \tag{1}$$

where $\mathbf{p}_i$ denotes the embedding of position i, *i.e.*, frame index. $\mathbf{r}_j$ denotes the embedding of landmark-type j. Both $\mathbf{p}_i$ and $\mathbf{r}_j$ are learned by supervised training.

3.2.2 Narrative Comment Graph - G_c. In the video titling task, the nouns and corresponding adjectives are of high importance in the narrative comment sentence. To capture these product-related words and their dependencies, we propose first to perform syntactic dependency parsing [43] on the comment sequence and build a narrative comment graph G_c accordingly. Words having dependencies will be connected.

3.2.3 Attributes Graph - G_a. Attributes are characteristics that define a particular product (such as appearance and functionality). Each video is associated with one product in Taobao, and the seller has manually annotated many key-value pairs, such as "Color:White", for the product. In practice, we keep the essential values as the attributes set. We represent the attributes set as fully-connected graph G_a.

3.2.4 Edge weights. The edges weights of all three graphs are obtained using the same schema, *i.e.*, a learned similarity function.

$$\mathbf{e}_{r,s} = \mathbf{W}_d[\mathbf{v}_r, \mathbf{v}_s] + \mathbf{b}_d \tag{2}$$

where $\mathbf{e}_{r,s}$ is a scalar, representing the edge weight of node v_r and v_s. $\mathbf{W}_d$ is a matrix denoting the similarity measurement and $\mathbf{b}_d$ is a bias term. A schematic graph representation example for these three kinds of graph is shown in Figure 2.

3.3 Granular-level Interaction Modeling

3.3.1 Information propagation. As a common practice, we leverage GNNs (Graph Neural Networks) as a trainable schema to aggregate information in local neighborhood nodes and exploit the rich information inherent in graph structure for each graph. Specifically, given the previous feature $\mathbf{v}_i$ of node v_i, the updated feature $\bar{\mathbf{v}}_i \in \mathbb{R}^{D_{out}}$ can be computed as:

$$\tilde{\mathbf{v}}_i^g = \sigma(\mathbf{W}_r^g \mathbf{v}_i + \mathbf{W}_n^g \,\mathrm{MAX}(\{\mathbf{e}_{i,j} * \mathbf{v}_j, j \in \mathcal{N}(i)\}) + \mathbf{b}_n^g) \quad (3)$$

$$\bar{\mathbf{v}}_i = \tilde{\mathbf{v}}_i^g * (\mathbf{W}_r^h \mathbf{v}_i + \mathbf{W}_n^h \,\mathrm{MAX}(\{\mathbf{e}_{i,j} * \mathbf{v}_j, j \in \mathcal{N}(i)\}) + \mathbf{b}_n^h) \quad (4)$$

where $\mathbf{W}_r^g, \mathbf{W}_n^g, \mathbf{W}_r^h, \mathbf{W}_n^h \in \mathbb{R}^{D_{out} \times D_{in}}$ are learnable linear transformations used to project the root feature and neighbors features into a joint space. $\mathbf{b}_n^g, \mathbf{b}_n^h \in \mathbb{R}^{D_{out}}$ are bias terms. D_{in} and D_{out} are the dimensions of the original node feature and the updated node feature. $\mathcal{N}(i)$ denotes the neighbors index set of node v_i. We adopt the element-wise max function MAX for the empirical effectiveness over the others, such as the element-wise average. σ is the element-wise sigmoid function, and $*$ denotes the element-wise multiplication. Equation 3 aims to obtain a gate vector $\tilde{\mathbf{v}}_i^g$, which is designed to control how much information is needed for each position in $\bar{\mathbf{v}}_i$ when updating. In our experiment, this design performs consistently better than the vanilla design without gate and GCNs [16], which is widely used in recent video graph modeling.

3.3.2 Global-local Aggregation. To capture the inter-actions across heterogeneous graphs in an end-to-end manner, we propose the *global-local aggregation* module, termed *GLA*. Given the query graph $G_q = \{v_i^q\}_{i=1,\ldots,N_q}$, and the context graph $G_e = \{v_i^e\}_{i=1,\ldots,N_e}$, we aim to obtain an aggregated graph G_{agg} with the same structure as G_q using two sub-modules, *i.e.*, *global gated access*, and *local attention.* We use $\mathbf{v}^q \in \mathbb{R}^{D_q}$ and $\mathbf{v}^e \in \mathbb{R}^{D_e}$ to denote the dense representation for the query node v^q and the context node v^e.

Global gated access We first obtain the global representation $\mathbf{v}^{q*}$ of the query graph by global-average pooling, and then encourage relevant information in each context node $\mathbf{v}^e$ and suppress irrelevant ones using a gate function. The query-aware context feature $\mathbf{v}_j^{eq} \in \mathbb{R}^{D_{eq}}$ can be obtained by:

$$\mathbf{v}^{q*} = 1/N_q \sum_{i=1}^{N_q} \mathbf{v}_i^q \quad (5)$$

$$\mathbf{g}_j = \sigma(\mathbf{W}_g[\mathbf{v}^{q*}, \mathbf{v}_j^e] + \mathbf{b}_g) \quad (6)$$

$$\mathbf{v}_j^{eq} = (1 - \mathbf{g}_j) * (\mathbf{W}_q \mathbf{v}^{q*} + \mathbf{b}_q) + \mathbf{g}_j * (\mathbf{W}_e \mathbf{v}_j^e + \mathbf{b}_e) \quad (7)$$

where $\mathbf{W}_g \in \mathbb{R}^{1 \times (D_q + D_e)}$, $\mathbf{W}_q \in \mathbb{R}^{D_{eq} \times D_q}$ and $\mathbf{W}_e \in \mathbb{R}^{D_{eq} \times D_e}$ are linear transformation matrices. $\mathbf{W}_g$ models the relevance between the global query graph representation $\mathbf{v}^{q*}$ and one context node $\mathbf{v}_j^e$. σ denotes the sigmoid function. $\mathbf{g}_j \in [0, 1]$ shows the relevance. $\mathbf{W}_q$ and $\mathbf{W}_e$ project the original representations into a query-context joint space (global). Intuitively, a larger $\mathbf{g}_j$ will encourage globally relevant information within the context graph. Smaller $\mathbf{g}_j$ will suppress the irrelevant and consider more global query information.

Local attention While the above process focuses on recognizing globally relevant information in the context graph (globally), the local attention sub-module aims to further filter important features related to each individual query node (locally). We perform node-level additive attention [1] which allows better selectivity in distilling relevant and necessary information. Specifically, we compute the final aggregated node vector $\mathbf{v}_i^{agg}$ as the following:

$$\mathbf{o}_{i,j} = \tanh(\mathbf{W}_o[\mathbf{v}_i^q, \mathbf{v}_j^{eq}] + \mathbf{b}_o) \quad (8)$$

$$\bar{\mathbf{o}}_{i,j} = \frac{\exp(\mathbf{W}_a \mathbf{o}_{i,j})}{\sum_k \exp(\mathbf{W}_a \mathbf{o}_{i,k})} \quad (9)$$

$$\mathbf{v}_i^{agg} = \mathbf{v}_i^q + \sum_{j=1}^{N_e} \bar{\mathbf{o}}_{i,j} * \mathbf{v}_j^{eq} \quad (10)$$

where $\mathbf{W}_o \in \mathbb{R}^{D_o \times (D_q + D_{eq})}$ and $\mathbf{W}_a \in \mathbb{R}^{1 \times D_o}$ jointly model the node-level relevance between one query node $\mathbf{v}_i^q$ and one updated context node $\mathbf{v}_j^{eq}$. $\bar{\mathbf{o}}_{i,j} \in [0, 1]$ indicates the relevance score. Finally, all locally relevant information have been distilled and summed up to obtain $\mathbf{v}^{agg}$.

3.3.3 Aggregating three graphs. We progressively perform information propagation for three individual graphs as well as the aggregated graphs, which are obtained by the global-local aggregation module, as depicted in Fig 3. In practice, we firstly aggregate the narrative comment graph and the attribute graph into the attribute-comment (AC) graph G_{ac}. We use the attributes graph as the query graph and distill relevant (globally and locally) information from the narrative comment graph due to the observation that attributes are always more systematic and meaningful. We finally obtain the video-attribute-comment (VAC) graph G_{vac} by viewing the video landmark graph as query and G_{ac} as the context graph since video can be the critical part for *video titling*. The VAC graph conveys the necessary granular-level cues, especially the landmark characteristics of the product, *e.g.*, "`V-collar`" or "`bat sleeve`".

3.4 Abstraction-level Story-line Summarization

The above *granular-level interaction modeling process* appropriately considers granular-level details of each product. As aforementioned, consumers usually customize their favorite products with the social surrounding and physical environment (*e.g.*, lighting, and decorations) in the demonstration, in order to learn the background-product interaction and generate the story-line video topic, we design the *abstraction-level story-line summarization* process.

Formally, given the aggregated graph G_{vac} and the frame-level features $\mathbf{X} = \{\mathbf{x}_t\}_{t=1,\ldots,N_v}$ (the detailed extraction process can be found in appendix A.3), we firstly use $\mathbf{X}$ as the query matrix, the nodes features of G_{vac}, *i.e.*, $\mathbf{V}_{vac}$, as context matrix and perform global-local aggregation (GLA) to obtain the aggregated feature sequence:

$$\bar{\mathbf{X}} = GLA(\mathbf{X}, \mathbf{V}_{vac}) \quad (11)$$

where *GLA* performs the same operations as illustrated in Section 3.3.2. This operation is intuitive under the empirical observation that the high-level picture should be consistent with the local design. Finally, we encapsulate the Gated Recurrent Unit (GRU) [7] to model the narrative structure of the frame feature sequence:

$$\tilde{\mathbf{x}}_t = [\bar{\mathbf{x}}_t, \mathbf{x}_t] \quad (12)$$

$$\mathbf{h}_t^S = \mathrm{GRU}(\mathbf{h}_{t-1}^S, \tilde{\mathbf{x}}_t) \quad (13)$$

We concatenate the globally-locally aggregated feature $\bar{\mathbf{x}}_t \in \bar{\mathbf{X}}$ and frame feature $\mathbf{x}_t$ as the input of GRU at time step t (The details of GRU can be found in the original paper[7]). We keep the hidden features $\mathbf{H}^S = \{\mathbf{h}_t^S\}_{t=1,\dots,N_v}$ of all steps for decoding.

3.5 Decoder

As a common practice, we employ the RNN as the decoder. Differently, we initialize the state $\mathbf{h}_0^D$ as the final state of the story-line summarization RNN, *i.e.*, $\mathbf{h}_0^D = \mathbf{h}_{N_v}^S$. In the decoding stage, we consider information within G_{vac}, G_{ac} and G_c graphs as well as the hidden features of the story-line reasoning RNN, $\mathbf{H}^S$. G_{vac} is obtained by using the visual landmark feature as a query and thus containing more visual information. Therefore, we consider graphs G_{ac} and G_c as necessary complements to incorporate the granular-level cues within the product attributes and the narrative comment. Formally, at decoder step m, we use the decoder hidden state $\mathbf{h}_m^D$ as query and apply the aforementioned local attention mechanism f_{la} to distill relevant information and generate titles depending on the four kinds of information, *i.e.*, G_{vac}, G_{ac}, G_c and $\mathbf{H}^S$:

$$\mathbf{x}_m^D = [f_{la}(\mathbf{h}_m^D, \mathbf{V}_{vac}), f_{la}(\mathbf{h}_m^D, \mathbf{V}_{ac}), f_{la}(\mathbf{h}_m^D, \mathbf{V}_c), f_{la}(\mathbf{h}_m^D, \mathbf{H}^S)] \tag{14}$$

Where $\mathbf{V}$ denotes all nodes features of a specific graph. We further concatenate the context feature $\mathbf{x}_m^D$ and the embedding $\hat{\mathbf{c}}_m$ of previously predicted word $\hat{c}_m$ as the input of decoding RNN. Specifically, we use GRU for its efficiency:

$$\mathbf{h}_{m+1}^D = \mathrm{GRU}(\mathbf{h}_m^D, [\hat{\mathbf{c}}_m, \mathbf{x}_m^D]) \tag{15}$$

3.6 Learning Objectives

Cross-Entropy Loss Our model minimizes the cross-entropy loss at each decoder step and we apply teacher forcing [41] during training. With the reference video title denoted as $C^R = \{c_m^R\}_{m=1,\dots,N_c}$, the cross-entropy loss can be defined as:

$$\mathcal{L}_{ce} = -\frac{1}{N_c} \sum_{m=1}^{N_c} \log f_P\left(c_m^R\right) \tag{16}$$

where $f_P\left(c_m^R\right)$ is the predicted probability of word c_m^R at timestamp m. The predicted probability distribution $P_m^O \in \mathbb{R}^{N_O}$ on the whole vocabulary O is obtained by applying one fully-connected layer $\mathbf{W}_p$ (with bias term $\mathbf{b}_p$) on the decoder hidden vector $\mathbf{h}_m^D$, followed by a softmax layer.

$$P_m^O = \mathrm{softmax}(\mathbf{W}_p \mathbf{h}_m^D + \mathbf{b}_p) \tag{17}$$

Generation Coverage Loss Despite the commonly used maximum-likelihood cross-entropy loss, we incorporate a probability-based loss similar to the coverage loss used in Text Summarization [32]. Instead of penalizing repetitively attending to the same location based on accumulative attention weights in the coverage loss, we directly suppress the repetitive generated words based on accumulative probability distribution. This inductive bias is reasonable due to the findings that repetitive words are mostly not expressive and attractive for *video titling*. Technically, we keep a generation coverage vector ς to store the accumulative probability distribution:

$$\varsigma_m = \sum_{m'=1}^{m-1} P_{m'}^O \tag{18}$$

To directly penalize the repetitive predicted words, our model minimizes the following loss:

$$\mathcal{L}_{gc} = \sum_m \sum_i \min\left(\varsigma_{m,i}, P_{m,i}^O\right) \tag{19}$$

Intuitively, when some word is predicted with high probability in the past, the corresponding value, *i.e.*, $\varsigma_{m,i}$, will be large. If the word probability in the the current prediction, $P_{m,i}^O$, is also large, the loss will become large. The final loss can be written as:

$$\mathcal{L} = \mathcal{L}_{ce} + \lambda_{gc}\mathcal{L}_{gc} \tag{20}$$

4 EXPERIMENTS

4.1 Data Collection

We collect a large-scale industrial *video titling* dataset, named *Taobao video titling dataset* (**T-VTD**) from real-world e-commerce scenario. T-VTD are comprised of 90,000 <video, comment, attributes, title> quadruples and each contains a product-oriented video stream from Mobile Taobao, a narrative comment uploaded alongside the video, human-nameable attributes from the associated product and a concise human-written video title. The basic statistics and comparison with benchmark video captioning (or related) datasets can be found in Appendix A.2.

4.2 Evaluation Metrics

To obtain a fair and comprehensive comparison between our methods and other state-of-the-arts, we evaluate our method in terms of various metrics, including reference-based metrics and retrieval-based metrics.

Reference-based metrics Following previous works in the field of video captioning, we capitalize on four reference-based metrics, BLEU4 [25], METEOR [2], ROUGE_L [19], CIDEr [35], to quantify the 4-gram precision, the stemming and synonymy matching, the longest common sequence overlap and the human-like consensus between the generated video titles and the references, respectively.

Retrieval-based metrics Despite the well-known advantages of the reference-based metrics, they are arguably not sufficient for our task because of the potentially large number of plausible solutions and the only-one reference for measurement. To this end, we adopt the retrieval-based evaluation protocol proposed by Abhishek *et al.*[8], which returns a sort of candidate titles based on the log-likelihood score rather than directly compare the generated title with the reference. In a similar spirit, for each test sample, we select candidate titles, including the title of the test sample, 50 plausible titles, 20 popular titles, and 29 randomly sampled titles. The plausibility is measured by the angular similarity of tf-idf dense representations of attributes sets. For popularity, since there are few repetitive titles in our dataset unlike in [8], we choose the top 20 titles with the most number of similar neighbors. We also use angular similarity, which distinguishes almost identical vectors much better, of tf-idf dense representations.

	T-VTD				Other Categories			
Method	BLEU-4	METEOR	ROUGE_L	CIDEr	BLEU-4	METEOR	ROUGE_L	CIDEr
M-MPLSTM	0.95	10.43	15.47	32.65	1.17	7.58	13.12	25.75
M-S2VT	1.52	11.64	18.03	41.74	1.70	9.49	17.26	39.35
M-HRNE	1.70	12.11	18.85	44.53	1.66	9.40	17.10	39.36
M-SALSTM	2.01	12.95	19.81	48.63	**2.33**	10.65	18.42	47.01
M-RecNet	2.00	12.87	20.33	50.01	2.24	10.62	19.18	47.81
M-LiveBot	1.86	12.65	19.56	47.02	1.86	9.89	16.93	40.81
Gavotte	**2.28**	**13.58**	**21.33**	**54.14**	2.27	**10.99**	**19.38**	**49.13**

Table 2: Performance of gavotte and various re-implemented state-of-the-art methods (for a fair comparison) in terms of four frequently used reference-based metrics. We conduct experiments on the released T-VTD (left) and one internal dataset without comprising clothes samples (right).

Models	R@1 ↑	R@5 ↑	R@10 ↑	MR ↓	MRR ↑
M-MPLSTM	11.1	34.6	46.3	23.48	0.230
M-S2VT	21.3	52.1	67.6	12.23	0.362
M-HRNE	22.6	54.0	68.3	11.98	0.375
M-SALSTM	25.9	54.5	67.84	12.16	0.397
M-RecNet	26.6	56.4	69.1	12.04	0.408
M-LiveBot	19.4	44.8	58.0	17.08	0.320
Gavotte	**26.8**	**58.2**	**71.0**	**11.614**	**0.415**

Table 3: Performance analysis using the retrieval-based metrics on T-VTD.

4.3 Competitors

To the best of our knowledge, there are no methods doing precisely the same task as ours before. We re-implement and make necessary modifications (for a fair comparison) to various publicly available video captioning methods (or related) as our competitors. We mainly add separate encoders for additional inputs, *i.e.*, the human-nameable attributes, and narrative comment, and concatenate the encoder outputs. In detail, we adopt:

- *M-MPLSTM.* M-MPLSTM adds two average-pooling encoders over MPLSTM [37].
- *M-S2VT.* M-S2VT replaces the naive mean-pooling strategy in M-MPLSTM by LSTM encoders.
- *M-HRNE* We re-implement the HRNE model [24] by modeling each input modality with a separate hierarchical encoder.
- *M-SALSTM* Based on M-S2VT, M-SALSTM further leverages the effective soft attention mechanism by attending to all three encoded feature sets in each decoding step.
- *M-RecNet* We re-implemented RecNet [38] over M-SALSTM by reconstructing the initial features of all inputs.
- *M-LiveBot* We modify the LiveBot [23] by adding an additional transformer encoder to model the attributes.

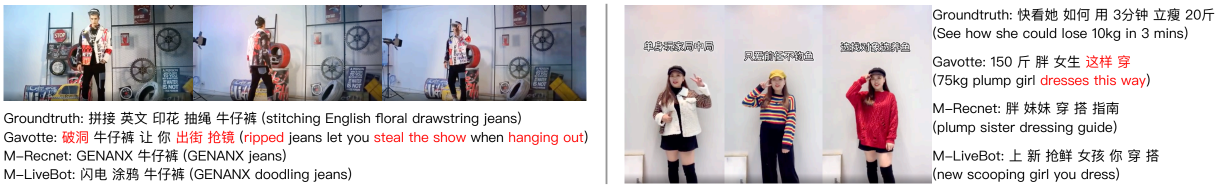
4.4 Performance Analysis

Quantitative Results In summary, the results on both the reference-based metrics and the retrieval-based metrics consistently indicate that our proposed method achieves better results against various Video Caption methods, including basic RNN models (M-MPLSTM and M-S2VT), attention-based RNN approach (M-SALSTM and M-RecNet) and transformer-based architecture (M-LiveBot).

Analogously to gains observed in other natural language generation tasks, the RNN based encoder (M-S2VT) is much better than the simple mean-pooling strategy (M-MPLSTM) in our task. A simple hierarchical design (M-HRNE) can lead to minor performance boost. Notably, attention-based architectures (M-SALSTM, M-RecNet, M-LiveBot, Gavotte) achieve the first-tier results due to its superior power on selecting important features and suppressing unnecessary ones. This ability is essential for our task since the three kinds of facts (*i.e.*, the consumer-generated video, narrative comment, and seller-generated attributes) can be noisy and unreliable in the real scenario. The reconstruction flow in M-RecNet guarantees that the decoded representation contains summarized information of the input and thus improving the M-SALSTM. The transformer-based M-LiveBot can not achieve competitive results than the M-SALSTM and M-RecNet. We relate the results to the inferior capacity of transformer-decoder than attentional RNN decoder [4] and its heavy dependence on pre-trained word embedding to avoid over-fitting. Our method outperforms the M-RecNet (the best among competitors) by almost 8.26% considering the CIDEr metric. We attribute the substantial improvements to the following reasons: 1) The graph representation can model special relationships such as the interactions of landmarks compared to the vanilla sequential dependency. 2) The proposed *global-local aggregation* module can better capture the inter-actions across heterogeneous graphs, resulting in an aligned and aggregated modeling. 3) The hierarchical design of granular-level interaction modeling and abstraction-level story-line summarization is essential for obtaining a comprehensive understanding of the video.

To verify the model capacity on non-clothes categories dataset, we conduct an additional experiment, of which the results are shown in Table 2. This internal dataset contains 90,000 videos belonging to three categories, *i.e.*, toys, makeups, and kitchen tools. Products of these categories do not have well pre-defined landmarks like clothes (*e.g.*, collar, and sleeve). We choose to first locate the presence of product with a bounding box and view the equally divided 3x3 areas of the bounding box as landmarks. We represent these nine landmarks using average-pooled features. This setting can be viewed as an approximation to the setting for clothes. The improvements over other competitors are comparatively smaller than the improvement on T-VTD. This is a reasonable result due to the coarser landmark feature extraction. Despite these disadvantages, our method achieves the best result on most measurements, which further demonstrate the merit of our design.

Human Evaluation To obtain a more reliable and precise measurement of the generation results, we conduct human judgments



Groundtruth: 拼接 英文 印花 抽绳 牛仔裤 (stitching English floral drawstring jeans)
Gavotte: 破洞 牛仔裤 让 你 出街 抢镜 (ripped jeans let you steal the show when hanging out)
M–Recnet: GENANX 牛仔裤 (GENANX jeans)
M–LiveBot: 闪电 涂鸦 牛仔裤 (GENANX doodling jeans)

Groundtruth: 快看她 如何 用 3分钟 立瘦 20斤
(See how she could lose 10kg in 3 mins)

Gavotte: 150 斤 胖 女生 这样 穿
(75kg plump girl dresses this way)

M–Recnet: 胖 妹妹 穿 搭 指南
(plump sister dressing guide)

M–LiveBot: 上 新 抢鲜 女孩 你 穿 搭
(new scooping girl you dress)

Figure 4: Title examples generated by Gavotte, M-RecNet, and M-Livebot with sampled frames of corresponding input videos. Results show that our model can generate more fluent and meaningful titles with attractive buzzwords like "style-freshing".

Models	Fluency	Diversity	Overall Quality
M-RecNet	3.671	3.497	3.310
M-LiveBot	3.901	3.613	3.407
Gavotte	**3.997**	**3.671**	**3.536**

Table 4: Human evaluation results across three aspects.

Models	BLEU-4	METEOR	ROUGE_L	CIDEr
Gavotte	2.28	13.58	21.33	54.14
- HSR	2.20	13.32	21.02	53.06
- GGA	2.13	13.24	20.89	52.57
- LA	2.06	13.33	20.61	51.48

Table 5: Model ablations by progressively removing three components.

concerning the following three aspects [5, 18]: Fluency, Diversity and Overall Quality. *Fluency* is designed to grade how fluent the generated titles are. The score of *Diversity* reflects whether the model generates refreshing content and is expected to be low when the titles are repetitive and dull. *Overall Quality* is intended to measure whether the titles are consistent with the three kinds of information, *i.e.*, video, comment, and attributes. The score range for each aspect is from 1 to 5. We consider M-RecNet and M-LiveBot as the representatives of RNN-based methods and transformer-based methods, and thus as competitors. We ask crowd workers in Taobao to grade the generation results of randomly sampled 1000 instances from the test set. Each instance contains the video link, the user-written comment, the attributes of the associated product, and generation results from three models. Workers are required to first watch the video before reading the other inputs and titles. According to the evaluation results shown in Table 4, Gavotte can generate more fluent and more diversified titles than the other two competitors. As for the measurement of faithfulness to inputs and some grounded facts (overall quality), Gavotte achieves a clear improvement over M-RecNet and M-LiveBot by +0.129 and +0.226 respectively. An interesting finding is that M-LiveBot performs much better than M-RecNet in human evaluation. By qualitatively comparing generation samples, we find that the titles generated by M-RecNet are short and accurate, which can be beneficial for automatic measurements. In contrast, a self-attention based model (M-Livebot) generates longer titles with more comparatively uncommon words, which can be more attractive for human judgers.

Qualitative Results Figure 4 shows some title generation samples from Gavotte and the other two competitors, *i.e.*, M-RecNet, and M-Livebot. Similar to the results in human evaluation, Gavotte can generate more fluent and attractive titles. Specifically, while the title of M-Recnet in the first case is less informative and the title of M-LiveBot in the second case is unfinished or broken, Gavotte generates smooth and meaningful title with the popular buzzwords – "steal the show". The results further demonstrate that Gavotte can recognize granular-level details like "ripped", clothes-level design like "jeans", frame-level product-background interaction effect like "steal the show", and video-level story-line topic like "dresses this way".

4.5 Ablation Studies

To better understand the behavior of different modules in our model, we surgically and progressively remove several components, including the *abstraction-level story-line summarization* process (HSR), the *gated global access* (GGC) and the *local attention* (LA) mechanism in global-local aggregation. The results are shown in Table 5. Overall, removing any of the components leads to a performance drop, which verifies the effectiveness of these components. Removing HSR will result in a failure to recognize the background, the overall style, and the story-line topic, and the metric scores decrease accordingly. During the graphs aggregation process, failing to access the global picture will result in a performance drop (from 53.06 to 52.57 considering the CIDEr metric). Notably, further removing the LA module means the total loss of the graph inter-action process. The obvious performance gap between the LA-removed model and the previous model demonstrates that the global-local aggregation module, especially the local attention sub-module, is an essential part for this task. In addition, this structure is similar to the M-SALSTM architecture except for the graph representation and GNN based encoder. The improvement over the M-SALSTM indicates the merit of leveraging both the flexible GNNs and granular-level cues within different kinds of information.

4.6 Online A/B Test

We deploy our model Gavotte on the *Guess You Like* scenario in mobile Taobao. Originally, the titles for recommended consumer-generated videos are generated as the truncated consumer-written comment. We conduct online tests on this baseline method and Gavotte under the framework of the A/B test. The testing set contains 10,000 video samples, which has no overlap with T-VTD. The count of total page views is around 400,000, and the number of all unique visitors is about 100,000. We keep the network traffic for both methods near the same. Gavotte improves the click-through-rate by +9.90% compared to the baseline. These results demonstrate that our model can generate meaningful and attractive titles, which further improve the performance of video recommendation.

5 CONCLUSION

In this paper, we propose a comprehensive information integration framework, named Gavotte, to generate descriptive and attractive titles for consumer-generated videos in e-commerce. We firstly represent three kinds of information (*i.e.*, the consumer-generated video, the narrative comment sentences, and the human-nameable attributes of the associated products) as graphs. Then we perform *granular-level landmark modeling* to exploit both the intra-actions within each graph using flexible GNNs and the inter-actions across graphs using proposed *global-local aggregation* module. This schema is designed to figure out the granular-level characteristics, such as the appearance of product landmarks, within different kinds of information, and aggregate them into a holistic representation. We further employ *abstraction-level story-line summarization* to capture the product-background interactions in the frame-level as well as the story-line topic in the video-level. As far as we know, this is the first piece of work that explores video graph modeling for general video captioning. We collect and release a corresponding large-scale dataset for reproduction and further research. The consistent quantitative improvement across various metrics reveal the effectiveness of the proposed method.

6 ACKNOWLEDGMENTS

The work is supported by the NSFC (61625107), Zhejiang Natural Science Foundation (LR19F020006), Fundamental Research Funds for the Central Universities (2020QNA5024, 2018AAA0101900), and a research fund supported by Alibaba.

REFERENCES

[1] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate.. In *ICLR*.

[2] Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*. 65–72.

[3] David L. Chen and William B. Dolan. 2011. Collecting Highly Parallel Data for Paraphrase Evaluation.. In *ACL*.

[4] Mia Xu Chen, Orhan Firat, Ankur Bapna, Melvin Johnson, Wolfgang Macherey, George Foster, Llion Jones, Mike Schuster, Noam Shazeer, Niki Parmar, and et al. 2018. The Best of Both Worlds: Combining Recent Advances in Neural Machine Translation.. In *ACL*.

[5] Qibin Chen, Junyang Lin, Yichang Zhang, Hongxia Yang, Jingren Zhou, and Jie Tang. 2019. Towards Knowledge-Based Personalized Product Description Generation in E-commerce.. In *KDD*.

[6] Guan-Yu Cheng, Hui-Qin Xiao, Jian Wei Li, Dong Wei Zhao, and Xiaosheng Wu. 2019. ICME Grand Challenge on Short Video Understanding.. In *ICME Workshop*.

[7] Kyunghyun Cho, Bart Van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder-decoder approaches. *arXiv preprint arXiv:1409.1259* (2014).

[8] Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual Dialog.. In *CVPR*.

[9] 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 *CVPR*.

[10] Christoph Feichtenhofer, Axel Pinz, and Richard P. Wildes. 2016. Spatiotemporal Residual Networks for Video Action Recognition.. In *NIPS*.

[11] Matthias Fey and Jan Eric Lenssen. 2019. Fast Graph Representation Learning with PyTorch Geometric.. In *ICLR Workshop*.

[12] Jonas Gehring, Michael Auli, David Grangier, Denis Yarats, and Yann N Dauphin. 2017. Convolutional sequence to sequence learning. In *ICML*.

[13] Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Subhashini Venugopalan, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2013. YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-Shot Recognition.. In *ICCV*.

[14] Kensho Hara, Hirokatsu Kataoka, and Yutaka Satoh. 2018. Can Spatiotemporal 3D CNNs Retrace the History of 2D CNNs and ImageNet?. In *CVPR*.

[15] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization.. In *ICML*.

[16] Thomas N Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *ICLR*.

[17] Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. 2002. Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions. *IJCV* (2002).

[18] Jiwei Li, Will Monroe, Tianlin Shi, Sébastien Jean, Alan Ritter, and Dan Jurafsky. 2017. Adversarial Learning for Neural Dialogue Generation.. In *EMNLP*.

[19] Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *ACL*. 74–81.

[20] Jingyuan Liu and Hong Lu. 2018. Deep Fashion Analysis with Feature Map Upsampling and Landmark-Driven Attention. In *ECCV*. Springer.

[21] Ziwei Liu, Sijie Yan, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2016. Fashion Landmark Detection in the Wild.. In *ECCV*.

[22] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NIPS*.

[23] Shuming Ma, Lei Cui, Damai Dai, Furu Wei, and Xu Sun. 2019. LiveBot: Generating Live Video Comments Based on Visual and Textual Contexts.. In *AAAI*.

[24] Pingbo Pan, Zhongwen Xu, Yi Yang, Fei Wu, and Yueting Zhuang. 2016. Hierarchical Recurrent Neural Encoder for Video Representation with Application to Captioning.. In *CVPR*.

[25] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *ACL*. Association for Computational Linguistics, 311–318.

[26] Adam Paszke, Sam Gross, Soumith Chintala, Gregory Chanan, Edward Yang, Zachary DeVito, Zeming Lin, Alban Desmaison, Luca Antiga, and Adam Lerer. 2017. Automatic differentiation in PyTorch. In *NIPS-W*.

[27] Xufeng Qian, Yueting Zhuang, Yimeng Li, Shaoning Xiao, Shiliang Pu, and Jun Xiao. 2019. Video Relation Detection with Spatio-Temporal Graph.. In *MM*.

[28] Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal. 2013. Grounding Action Descriptions in Videos. *TACL* (2013).

[29] Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent Multi-sentence Video Description with Variable Level of Detail.. In *GCPR*.

[30] Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for Movie Description.. In *CVPR*.

[31] Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele. 2013. Translating Video Content to Natural Language Descriptions.. In *ICCV*.

[32] Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get To The Point: Summarization with Pointer-Generator Networks.. In *ACL*.

[33] Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in Homes: Crowdsourcing Data Collection for Activity Understanding.. In *ECCV*.

[34] Atousa Torabi, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. 2015. Using Descriptive Video Services to Create a Large Data Source for Video Annotation Research. *Arxiv* (2015).

[35] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *CVPR*. 4566–4575.

[36] Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond J. Mooney, Trevor Darrell, and Kate Saenko. 2015. Sequence to Sequence - Video to Text.. In *ICCV*.

[37] Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond J. Mooney, and Kate Saenko. 2015. Translating Videos to Natural Language Using Deep Recurrent Neural Networks.. In *NAACL*.

[38] Bairui Wang, Lin Ma, Wei Zhang, and Wei Liu. 2018. Reconstruction Network for Video Captioning.. In *CVPR*.

[39] Wenguan Wang, Xiankai Lu, Jianbing Shen, David J Crandall, and Ling Shao. 2019. Zero-shot video object segmentation via attentive graph neural networks. In *ICCV*.

[40] Xiaolong Wang and Abhinav Gupta. 2018. Videos as Space-Time Region Graphs.. In *ECCV*.

[41] Ronald J. Williams and David Zipser. 1989. A Learning Algorithm for Continually Running Fully Recurrent Neural Networks. *Neural Computation* (1989).

[42] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher J. Pal, Hugo Larochelle, and Aaron C. Courville. 2015. Describing Videos by Exploiting Temporal Structure.. In *ICCV*.

[43] Daniel Zeman, Jan Hajic, Martin Popel, Martin Potthast, Milan Straka, Filip Ginter, Joakim Nivre, and Slav Petrov. 2018. CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies.. In *ACL*.

[44] Kuo-Hao Zeng, Tseng-Hung Chen, Juan Carlos Niebles, and Min Sun. 2016. Title Generation for User Generated Videos.. In *ECCV*.

[45] Runhao Zeng, Wenbing Huang, Mingkui Tan, Yu Rong, Peilin Zhao, Junzhou Huang, and Chuang Gan. 2019. Graph convolutional networks for temporal action localization. In *ICCV*.

A APPENDIX

A.1 Details on the Experimental Setup and Hyperparameters

Hyper-parameter Configuration The parameter setup of *granular level interaction modeling* process is shown in Figure 5. During training, the batch size is set to 64 and we use Adam optimizer [15] with the setting $\beta_1 = 0.9, \beta_2 = 0.999$, weightdecay $= 1 \times 10^{-4}$ and $\epsilon = 1 \times 10^{-8}$. Learning rate is 4×10^{-4}. We employ dropout rate of 0.2 and batch normalization after graph information propgation. We also applied a dropout rate of 0.5 for RNNs and linear layers as regularization. The hidden size of both story-line summarization RNN and decoder RNN is set to 512. The loss weight λ_{gc} is set to 0.1. At the inference stage, we use the greedy strategy to generate the final title.

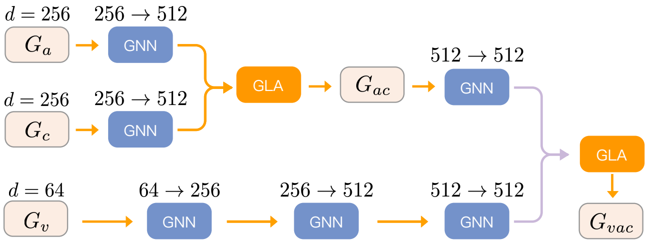
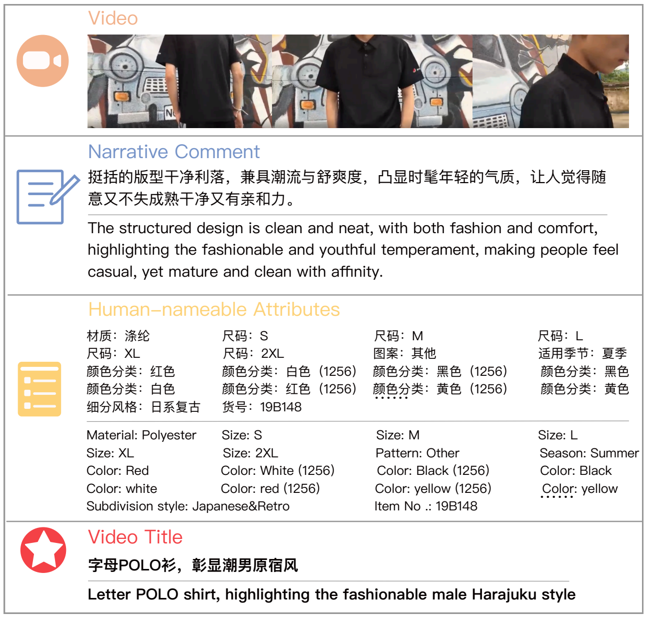


Figure 5: Architecture and parameter configuration of granular-level interaction modeling process.

Hardware & Software Configuration The experiments are conducted on a Linux server equipped with an Intel(R) Xeon(R) CPU E5-2690 v4 @ 2.60GHz, 512GB RAM and 1 NVIDIA Titan V GPU. All models are implemented in PyTorch [26] of version 1.2.0, pytorch-geometric [11] of version 1.3.2 and Python 3.6. We'll soon release the main code of Gavotte for reproduction and further development.

A.2 Additional Details on the Dataset

The summarized statistics of T-VTD and comparisons with other frequently used benchmark video captioning datasets are shown in Figure 6. Specifically, T-VTD contains 90000 videos with a total length of 755.73 hours, which is much larger than current datasets. We choose three best-sold categories, *i.e.*, men wear, women wear and children wear, and collect 30,000 videos for each category. As for natural language data, T-VTD has totally 3,878,436 words (in titles and comments) with a vocabulary of 68232. The nature of abundant vocabulary and little repetitive information poses a direct challenge to fully understand the semantic information and avoid obtaining high scores just by overfitting to some biases. In contrast to existing datasets which mostly concern the object-level information (such as the object categories and actions), sentences in our dataset incorporate rich details in four levels: 1) landmark characteristics in the granular level; 2) the overall appearance and functionality in the product level; 3) the interaction with the background in the frame level 4) the story-line topic in the video level. These distinguishing features of T-VTD pose new challenges from the industry scenario. We summary the statistics of video durations in our dataset in Table 8. The video durations in our dataset are mainly 15-30s, with an average video length around 30.23s. The longest video can reach 600 seconds (about 10 minutes). As shown in Table 7 and 9, the average length of elements in the titles, comments and attributes are 6.6, 36.49 and 17.54, respectively. The vocabulary size is 27,976 for video titles and 59,109 for narrative comments. There are 81 different attribute-key types and 266,648 different attribute-value types, *i.e.*, the vocabulary size of human-nameable attributes. A real case in our dataset is shown in Figure 6. It can be seen that the comment sentences mainly narrate the preference for different aspects of products. Although the attributes of associated products structurally specify the human-nameable qualities of the products, these attributes can be noisy since they may cover all possible choices (such as different colors) and not exactly the product in the video. Models are required to fully understand the video rather than simply distill information from the attributes or the comment.



挺括的版型干净利落，兼具潮流与舒爽度，凸显时髦年轻的气质，让人觉得随意又不失成熟干净又有亲和力。

The structured design is clean and neat, with both fashion and comfort, highlighting the fashionable and youthful temperament, making people feel casual, yet mature and clean with affinity.

材质：涤纶	尺码：S	尺码：M	尺码：L
尺码：XL	尺码：2XL	图案：其他	适用季节：夏季
颜色分类：红色	颜色分类：白色（1256）	颜色分类：黑色（1256）	颜色分类：黑色
颜色分类：白色	颜色分类：红色（1256）	颜色分类：黄色（1256）	颜色分类：黄色
细分风格：日系复古	货号：19B148		

Material: Polyester	Size: S	Size: M	Size: L
Size: XL	Size: 2XL	Pattern: Other	Season: Summer
Color: Red	Color: White (1256)	Color: Black (1256)	Color: Black
Color: white	Color: red (1256)	Color: yellow (1256)	Color: yellow
Subdivision style: Japanese&Retro		Item No .: 19B148	

字母POLO衫，彰显潮男原宿风

Letter POLO shirt, highlighting the fashionable male Harajuku style

Figure 6: A <video, comment, attributes, title> quadruple data sample in T-VTD.

A.3 Details on Data Pre-processing

For text pre-processing, we remove the punctuations and tokenize sentences using Jieba Chinese Tokenizer [2]. Our vocabulary contains all attributes values, comment tokens and ground-truth title tokens. Since the real-world text data is noised and many expressions can be confusing or meaningless, such as brands and homophonic words. We roughly filter them by replacing low-frequency tokens (less than 50) with the special token $< unk >$, resulting in 6347 tokens in total. The length limitations for title, comment and attributes are 12, 50 and 15, respectively. Text with number of tokens more than the corresponding limitation will be truncated. We add a special $< sos >$ token as the first word for the title and a $< eos >$ at the

[2] https://github.com/fxsjy/jieba

Dataset	Context	#Video	#Sentence	#Word	#Vocabulary	Total Duration(hrs)
MSVD [3]	multi-category	1,970	70,028	607,339	13,010	5.3
YouCook [9]	cooking	88	2,668	42,457	2,711	2.3
TACos [28]	cooking	123	18,227	146,771	28,292	15.9
TACos M-L [29]	cooking	185	14,105	52,593	-	27.1
MPII-MD [30]	movie	94	68,375	653,467	24,549	73.6
M-VAD [34]	movie	92	55,905	519,933	18,269	84.6
VTW [34]	multi-category	18,100	44,603	-	23,059	213.2
MSR-VTT [44]	multi-category	7,180	200,000	1,856,523	29,316	41.2
Charades [33]	human	9,848	27,847	-	-	82.01
T-VTD	e-commerce	90,000	180,000	3,878,436	68,232	755.73

Table 6: Comparison between T-VTD with benchmark video captioning datasets, considering various capacity indicators..

	avg_len	total_len	vocab
title	6.6	594,279	27,976
comment	36.49	3,284,157	59,109

Table 7: Basic statistics of titles and comments in T-VTD.

	average	min	Q1	median	Q3	max
video duration (s)	30.23	1.5	15.72	23.56	32.2	600.08

Table 8: Statistics of the video duration. (Q1 denotes the lower quartile and Q3 denotes the upper quartile.)

	avg_num	total_num of keys	total_num of values	vocab of keys	vocab of values
attributes	17.54	1,578,777	2,604,904	81	266,648

Table 9: Basic statistics of attributes in T-VTD.

end. When the $< sos >$ token is predicted in the decoding stage, the generation will be terminated.

For video processing, we first uniformly sample 30 frames per video. For landmark feature extraction, we extract the product area using internal product detector for all sampled frames. Then we use the pre-trained landmark detector [3] provided by [20]. Specifically, the backbone model $VGG16$ takes each frame as input and output the activations of shape $512 \times 7 \times 7$ from layer *pooled_5*. This feature map is forwarded to the landmark decoder, which produce the landmark-oriented features of shape $64 \times 14 \times 14$ and the mask-like landmark maps of shape $8 \times 56 \times 56$, *i.e.*, 8 landmarks maps of shape $\times 56 \times 56$ each. We downsampled each landmark map to have the same width and height as the landmark-oriented features. Deriving from the observation that intermediate feature map and emergent patterns are highly correlated, we normalize each landmark map as weights using softmax and compute weighted sum over the landmark-oriented features as the landmark feature.

For frame-level feature extraction, we use the same model as landmark feature extraction and obtain the activations of shape $128 \times 7 \times 7$ from layer *conv4* for each frame. We use the global average pooling result as the frame feature.

The dataset we use for training is a subset (84394 samples) of the released dataset (90000 samples) due to the the data pre-processing (mainly the low-frequency words removal procedure) after which many words in the comment and elements in the attributes set will be replaced by $< unk >$. Specifically, we remove the following 3 kinds of samples: 1) sample with less than 2 non-$< unk >$ elements (*i.e.*, where will be no edges in the graph) in the attributes set. 2) sample with 0 non-$< unk >$ words in the title. 3) sample with less than 11 non-$< unk >$ nodes in or less than 5 edges in the narrative comment graph. Overall, we mainly remove samples with little information within either one kind of fact (narrative comment or human-nameable attributes) or the ground truth title to make the data more reliable.

We randomly split the whole dataset by train 65%, validation 5% and test 30%, resulting in 54856 samples for training, 4220 samples for validation and 25318 samples for testing.

[3] https://github.com/fdjingyuan/Deep-Fashion-Analysis-ECCV2018