TagGPT: Large Language Models are Zero-shot Multimodal Taggers

Chen Li[1] **Yixiao Ge**[1] **Jiayong Mao**[2] **Dian Li**[2] **Ying Shan**[1]

[1]ARC Lab, Tencent PCG
[2]Foundation Technology Center, Tencent PCG

Abstract

Tags are pivotal in facilitating the effective distribution of multimedia content in various applications in the contemporary Internet era, such as search engines and recommendation systems. Recently, large language models (LLMs) have demonstrated impressive capabilities across a wide range of tasks. In this work, we propose TagGPT, a fully automated system capable of tag extraction and multimodal tagging in a completely zero-shot fashion. Our core insight is that, through elaborate prompt engineering, LLMs are able to extract and reason about proper tags given textual clues of multimodal data, *e.g.*, OCR, ASR, title, *etc*. Specifically, to automatically build a high-quality tag set that reflects user intent and interests for a specific application, TagGPT predicts large-scale candidate tags from a series of raw data via prompting LLMs, filtered with frequency and semantics. Given a new entity that needs tagging for distribution, TagGPT introduces two alternative options for zero-shot tagging, *i.e.*, a generative method with late semantic matching with the tag set, and another selective method with early matching in prompts. It is well noticed that TagGPT provides a system-level solution based on a modular framework equipped with a pre-trained LLM (GPT-3.5 used here) and a sentence embedding model (SimCSE used here), which can be seamlessly replaced with any more advanced one you want. TagGPT is applicable for various modalities of data in modern social media and showcases strong generalization ability to a wide range of applications. We evaluate TagGPT on publicly available datasets, *i.e.*, Kuaishou and Food.com, and demonstrate the effectiveness of TagGPT compared to existing hashtags and off-the-shelf taggers.[1]

1 Introduction

Tags, as concise descriptions of semantic content, have been demonstrated to play an incredibly important role in numerous downstream tasks [8]. They facilitate the system's quick and accurate understanding of user interests and search intentions, simplifying the search [9, 6] and recommendation [16, 1, 15] process as a result. For example, as illustrated in Fig. 1, on social media platforms with multimodal content, users are permitted to use customized tags to effortlessly describe their published content, luring users with similar interests. Therefore, obtaining top-notch tags and constructing comprehensive tagging systems and corresponding taggers have always been important ways to improve the service quality of such applications.

The traditional sources for obtaining tags are primarily human annotation and language model extraction. However, tags obtained through manual annotation tend to be subjective and expensive [16], while those extracted by language models are constrained by limited understanding capabilities

[1]Project page: https://github.com/TencentARC/TagGPT

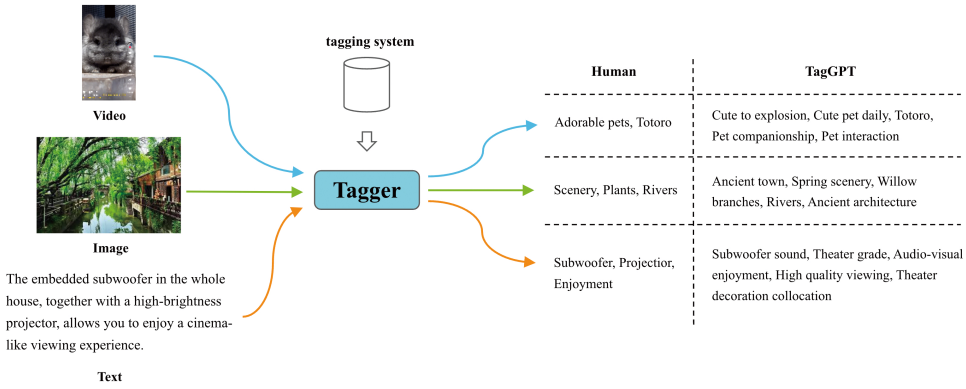


Figure 1: Given multimodal content from social media (*e.g.*, Twitter, Weibo, *etc.*), a tagger aims to produce several phrases that can properly describe the content and reflect the user's interests.

and may lack reliability [8]. Especially with the rapid expansion of the semantic space in today's social networks and the increasing diversity of data modalities, the challenge of quickly obtaining high-quality tags from massive multi-modal data has become a pressing research and application issue.

To construct a high-quality content-oriented tagging system and corresponding zero-shot tagger, we propose `TagGPT`, a generative solution in this paper. Our approach draws inspiration from the impressive inference capabilities of Large Language Models (LLMs) and their exceptional performance in zero-shot settings [3; 13; 19; 17; 5]. We begin by converting multimodal data into unified textual clues through mature unimodal models. These clues are then inserted into a discrete instruction prompt, which feeds into an LLM to generate potential tags. By collecting tags from large-scale inputs, we can collate a set of content tags. We then utilize post-processing techniques such as unsupervised semantic similarity fusion and frequency distribution filtering to obtain a high-quality content-oriented tagging system. Finally, we design two distinct zero-shot taggers that rely on the generation or context inference abilities of LLMs for zero-shot data tagging.

We have built benchmarks using data from two popular applications and conducted experiments. Our experimental results have shown that `TagGPT` has the ability to create a top-notch tagging system and allocate suitable tags to the input multimodal data in the zero-shot scenario. For the complete project code and additional examples, please refer to `https://github.com/TencentARC/TagGPT`.

2 TagGPT

`TagGPT` is a framework designed for processing multimodal content. The framework boasts a pipeline that automatically constructs a tagging system, along with a tagger that completes zero-shot annotation. At the heart of this framework lies the ability to leverage the capability of LLMs to obtain a high-performance, top-quality tagging system, all at a low cost. In the following sections, we will introduce the two primary modules that comprise `TagGPT`.

2.1 Tagging System Construction

The construction of a content-oriented tagging system aims to capture valuable semantic information from large-scale data and form a complete tagging system. In `TagGPT`, to be able to mine high-quality tags from large-scale multimodal data, as shown in Fig. 2, we propose a pipeline consisting of three sub-modules.

2.1.1 Textual clue converter

With the advancement of mobile networks and multimedia technology, popular applications like social networks, e-commerce platforms, and media information platforms are increasingly utilizing multimodal data as the primary information distribution carrier. The rapid development of a sound

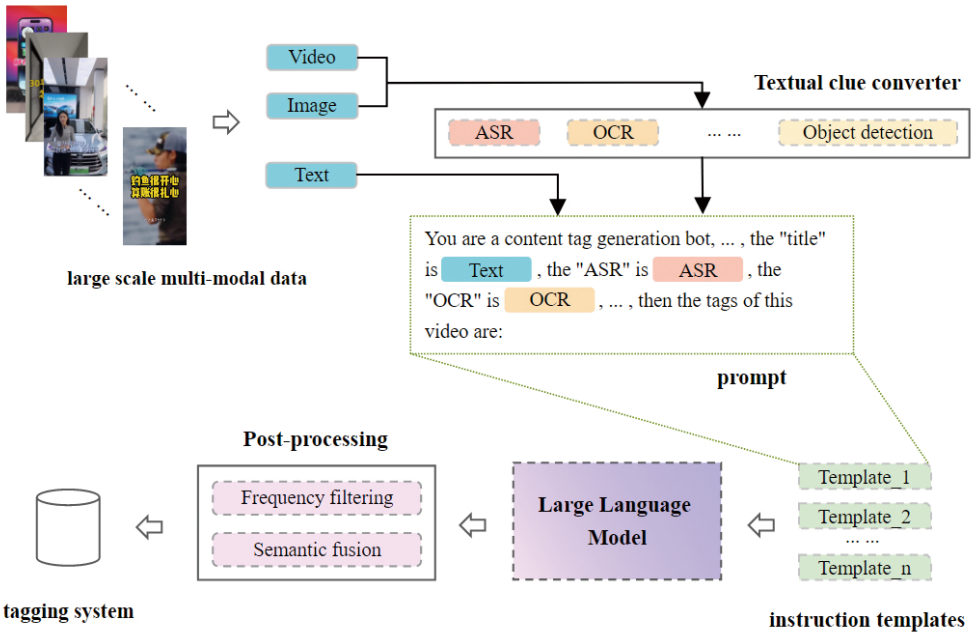


Figure 2: Given a series of raw data from a specific application, TagGPT is capable of building a high-quality tagging system in an entirely zero-shot manner without extra knowledge or human annotation. Such a paradigm enables instant tagging of new applications with zero labor cost.

tagging system based on such multimodal data can provide fundamental support for services such as search and recommendation.

Recently, LLM has performed well in content understanding, reasoning, and generation tasks, thus we attempt to apply LLM to the construction of tagging systems. To maximize the benefits of LLMs, TagGPT inspired by [22; 18] first processes multimodal data and extracts rich textual clues from it with the help of various unimodal basic models as input to downstream modules. For example, when dealing with video data, TagGPT employs robust Automatic Speech Recognition (ASR) [21] and Optical Character Recognition (OCR) [4] models to extract textual information from video frames and audio. Similarly, to meet the construction requirements of the tagging system in various domains, processing technologies such as object detection [2] and action recognition [10] can be utilized to provide textual clues from different perspectives for TagGPT.

2.1.2 Tag generation based on LLM

With the ever-evolving progress of LLMs in terms of training data scale, parameter scale, and training strategies, LLMs such as GPT [3; 13] and PaLM [5] have exhibited remarkable understanding, reasoning, and generation capabilities. Consequently, numerous downstream tasks have benefited from these models, achieving significant performance improvements. Recently, the introduction of InstructGPT [14] has further fortified the ability of LLMs to comprehend various task instructions, thus enhancing their capacity to transition between different tasks. In TagGPT, we strive to incorporate the formidable instruction understanding ability of LLMs into new tasks, as well as their potential to understand, reason, and generate content. This will facilitate the creation of a tagging system. More specifically, we have created multiple instruction templates for acquiring tags and filled in the corresponding positions with textual clues that correspond to the single multimodal data provided from upstream. The input text is then fed into the LLM, and the corresponding tag result can be obtained. After a large batch of generation operations, large-scale tags generated by LLMs can be obtained, resulting in a complete and comprehensive candidate tag set.

2.1.3 Post-processing

Of course, the tag set constructed in this manner is bound to contain a lot of noise and redundant semantics. Thus, TagGPT includes a post-processing module at the end of the tagging system construction in hopes of improving the quality of the tagging system through a series of operations. Firstly, the module truncates tags that are too high or too low based on their frequency; tags with a high frequency are likely to come from popularity deviation and lack sample distinguishability [23], while tags with low frequency may come from very few samples and lack significant tag value [20]. The module then checks the semantic similarity of the remaining tags and fuses those with similar semantics to further reduce the tagging system's scale. Specifically, all tags in the tagging system are encoded based on an unsupervised pre-trained text encoder to obtain their corresponding representations. The module then calculates the cosine similarity to determine whether the tags should be fused.

So far, we can automatically construct a corresponding tagging system based on given large-scale multimodal data.

2.2 Zero-shot Multimodal Tagger

In addition to automatically constructing a content-oriented tagging system, TagGPT also provides corresponding zero-shot taggers. As shown in Fig. 3, according to the logic of annotation, we will introduce two different taggers separately.

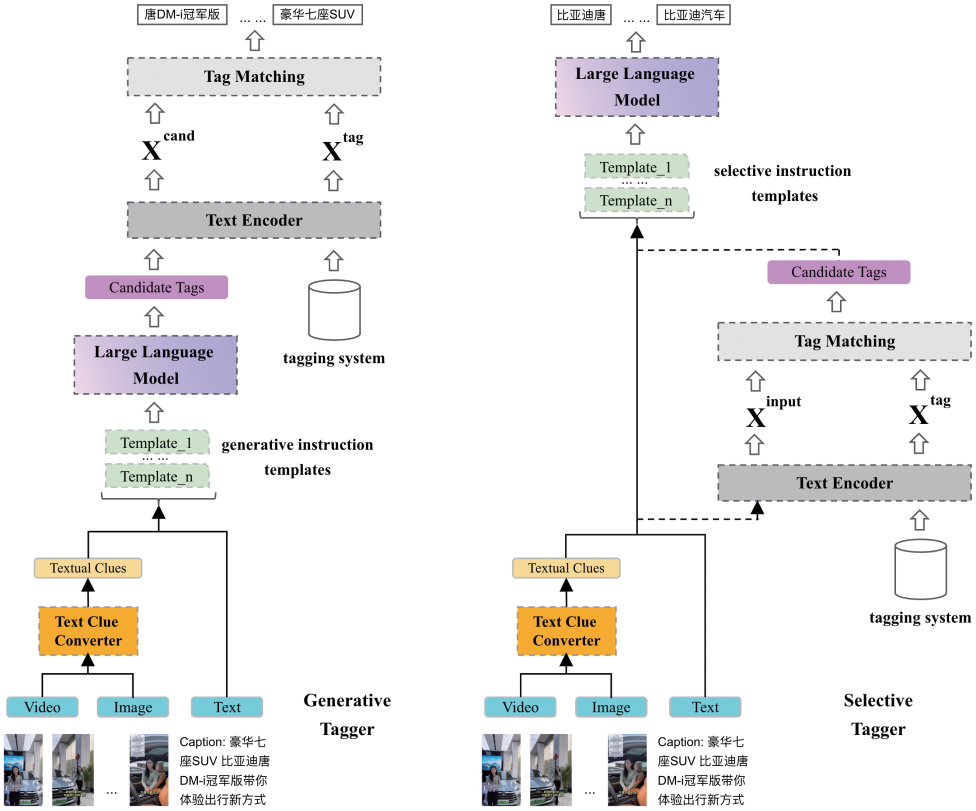


Figure 3: Given the tagging system established in Figure 2, TagGPT enables zero-shot tagging of new data in two alternative paradigms.

2.2.1 Generative tagger

Due to its powerful understanding and reasoning abilities, LLM is an excellent zero-shot tagger. The previous tagging system was constructed based on this capability. Consequently, if the tags inferred by LLM can be quickly matched with the existing tagging system for semantic similarity, then zero-shot tagging can be achieved. Building on this idea, we propose a generative tagger that

converts input multimodal samples to textual clues, fills in the corresponding positions of candidate instruction templates, and uses LLM to provide candidate tags. Candidate tags are encoded as $\mathbf{X}^{cand} \in \mathbf{R}^{n \times m}$ using a pre-trained text encoder, while tags in existing tagging systems are encoded as $\mathbf{X}^{tag} \in \mathbf{R}^{N \times m}$, n, N, m denote the number of candidate tags, the number of tags in the tagging system and the representation dimension. The matching score matrix $\mathbf{S} \in \mathbf{R}^{N \times n}$ is then calculated as $\mathbf{S}_{i,j} = \mathbf{X}_i^{tag} \cdot \mathbf{X}_j^{cand\top} / (|\mathbf{X}_i^{tag}| \times |\mathbf{X}_j^{cand}|)$. Finally, the matching result can be directly selected through the preset threshold. The left part of Fig. 3 illustrates this process.

2.2.2 Selective tagger

When LLM is put into practice, people discover its strong ability to understand context, and some methods even use in-context learning strategies to guide LLM in producing more reasonable outcomes [11]. This inspired us to develop another zero-shot tagging paradigm, as depicted in the right part of Fig. 3. Specifically, we first convert the given multimodal samples into textual clues. Unlike the generative tagger, we input it along with the tags in the tagging system into a pre-trained text encoder to obtain the corresponding vector representation, *i.e.*, $\mathbf{X}^{input}$ and $\mathbf{X}^{tag}$. By matching similarities between the text and tags, we can obtain a series of related candidate tag sets. Finally, by incorporating the original textual clues and candidate tag sets into the selective instruction template and feeding them into LLM, we can obtain the final annotation result for a given multimodal sample.

3 Experiments

In this section, we will test `TagGPT`'s ability to construct the tagging system and assign tags to the given multimodal data in the zero-shot setting on two real-world social network datasets.

3.1 Dataset and Metrics

Dataset	# items	caption	category	OCR	ASR
Kuaishou	222k	✔	✔	✔	✔
Food.com	5.4k	✔	✔	✗	✔

Table 1: The statistics of the datasets.

We have chosen two renowned websites, Kuaishou[2] [12] and Food.com[3], as the data sources for our experiment. The detailed statistics are presented in Tab. 1.

When evaluating the `TagGPT`-based tagging system, we will use the source's own hashtag system (*i.e.*, extracted from Kuaishou and food.com) as the baseline. To comprehensively evaluate the tagging system's quality, we will refer to previous research designs and consider the following metrics:

- **Popularity**: When a tag is consistently assigned to a group of related objects by the majority of users, it indicates that the tag holds stable value and reduces the likelihood of it being considered spam. However, if the tag is assigned too broadly (*i.e.*, too popular), its ability to distinguish between objects is greatly diminished, which defeats the purpose of the tag. To assess the indicator, we intend to track the distribution of tags for both the baseline and the `TagGPT`-based tagging system using identical data. This will enable us to observe the deviation in tag popularity between the two tagging systems.
- **Practicality**: This metric includes two sub meanings, *i.e. least effort* and *high coverage of multiple facets*. This pair of metrics are related, with the first one (*i.e.*, least effort) focusing on the number of tags assigned to a given object, while the second one focuses on the facet coverage of a given object. In fact, these two indicators are theoretically positively correlated, that is, when fewer tags are assigned, the number of facets they can cover decreases. In this paper, we intend to evaluate this pair of indicators by counting the number of tags in each given data under different tagging systems and assessing the semantic redundancy of

[2]https://www.kuaishou.com/
[3]https://www.food.com/

tags within each given data. Among them, the semantic redundancy of tags comes from the proportion of similar semantics in the corresponding tags for each given data.

- **Uniformity**: Tags may diverge due to personal habits and application scenarios, resulting in different descriptions of the same concept. While keeping these contents can improve the recall rate of annotations and enhance the user experience, excessive tags will increase the tagging system size and introduce noise. When evaluating, we will consider the semantic redundancy (*i.e.*, semantic similarity) between any two tags in the whole tagging system as the calculation method.

3.2 Setup

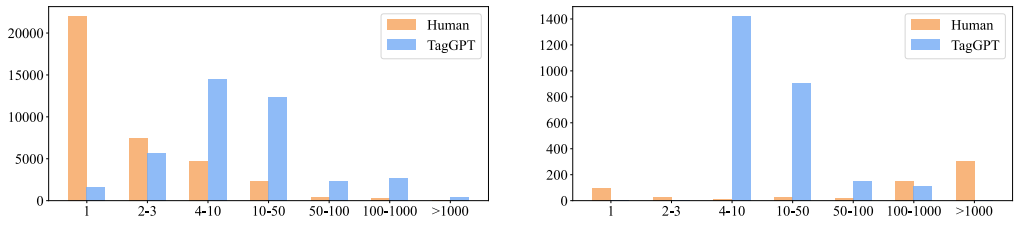
In the selection of LLM, we chose GPT 3.5 [14] (API interface, the version is *gpt-3.5-turbo*). The model of choice for the unsupervised text encoder is SimCSE [7]. In the judgment of semantic redundancy, we take the cosine similarity of the unsupervised semantic vector representation as the measure index. When the similarity is greater than or equal to 0.8, we consider that the two input objects have non-negligible semantic redundancy.

To be able to set a reasonable baseline for comparison for `TagGPT`, we take the tags provided by users when they upload videos for themselves as the baseline. Specifically, we extract the content of the given data spaced by "#" as the result of the user tagger and collect them as the tagging system of the user annotation. In addition, we also select two well-known visual[4] and text[5] online tagging methods for comparison.

3.3 Main Result

In this section, we will evaluate the performance of different sub-modules in `TagGPT` from the two aspects of the tagging system and taggers respectively.

3.3.1 Tagging system



(a) Kuaishou (b) Food.com

Figure 4: Statistical results of the metric “popularity” in the datasets. The horizontal axis denotes the number of times a tag is assigned to the data, and the vertical axis denotes the number of tags.

Fig 4 presents the statistical results of `TagGPT`'s tag distribution on the “popularity” metric. It is evident that the distribution of tags annotated by users is not optimal in either of the datasets. In the Kuaishou dataset, each tag is used by the user less than four times, and a staggering 60% of the tags are only used once. On the other hand, in the Food.com dataset, the tag distribution tends to be polarized, where some tags are used too infrequently while others are used too frequently. As mentioned earlier, if the tag popularity is too low, the system will have to face a large number of useless tags, making it difficult for users to efficiently match the correct information. Conversely, if the tag popularity is too high, the system will struggle to effectively distinguish data, leading to a suboptimal user experience. `TagGPT` exhibits excellent tag coverage in both datasets, with a significant number of tags being reused more than 10 times without being overused. This feature

[4] Azure Cognitive Services for Vision (ACSV): https://portal.vision.cognitive.azure.com/demo/generic-image-tagging.

[5] Hanlp: https://www.hanlp.com/demonstrate.html.

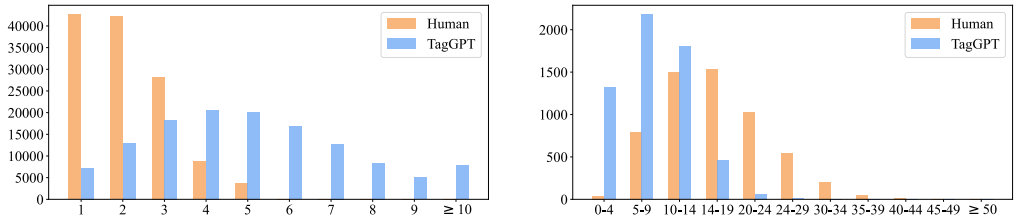
Dataset	Source	# tags	Redundancy Ratio
Kuaishou	Human	37,320	2.85%
	TagGPT	40,238	0.0%
Food.com	Human	458	0.97%
	TagGPT	2,598	0.0%

Table 2: “Uniformity” scores in the corresponding tagging systems for both datasets.

Dataset	selective tagger			generative tagger		
	Precision	Recall	F1	Precision	Recall	F1
Kuaishou	78.3	69.4	73.6	80.1	72.1	75.9
Food.com	81.5	75.0	78.1	83.7	76.9	78.7

Table 3: Quantitative results of two alternative taggers in TagGPT.

facilitates downstream recommendation, search, and other algorithms, ultimately enhancing their performance.



(a) Kuaishou (b) Food.com

Figure 5: Statistical results of “least effort” in the metric “practicality" in the dataset. The horizontal axis denotes the number of tags assigned to a single data sample, and the vertical axis is the number of samples.

To assess the metric “practicality” of tagging systems, we will analyze it from two angles: the number of tags assigned to each sample (i.e., "least effort") and the redundancy of tags within each sample (*i.e.*, “high coverage of multiple facets”). Fig 5 displays the tag count for the dataset, revealing that TagGPT assigns significantly more tags to each video than the baseline method. Furthermore, we calculated the average tag redundancy rate within each sample for both tagging systems. TagGPT’s redundancy rate is 0% in both datasets because it filters out redundant tags during tag statistics. In comparison, the baseline method sets a similarity threshold of 0.8 for statistics, resulting in an average tag redundancy of 5.86% and 3.77%. This demonstrates that TagGPT not only assigns more tags to each sample than the baseline method but also ensures that the semantic coherence between these tags is low, aligning with the diverse perspectives of each sample.

When calculating the “uniformity” score, we will pair all tags in the tagging system to determine their semantic similarity. We then calculate the proportion of tag pairs whose similarity exceeds the threshold of 0.8 among all tag pairs. The results are presented in Tab. 2, showing that the internal uniformity of the tagging system constructed by TagGPT is significantly better than the tagging system based on human annotation. This proves that TagGPT has lower tag redundancy. Furthermore, when considering the number of tags and their redundancy, TagGPT’s tags can cover a wider range of content perspectives.

(a) High-frequency tags inferred from the Kuaishou dataset.

(b) High-frequency tags inferred from the Food.com dataset.

Figure 6: The word cloud of the top 150 words in the tagging system constructed by TagGPT in both datasets.

3.3.2 Taggers

In this paper, we introduce two distinct zero-shot taggers, namely generative tagger and selective tagger. To assess the effectiveness of the taggers, we randomly set aside 100 samples from the dataset as a test set, which were not involved in the development of the tagging system. The performance of the taggers was evaluated through manual assessment, and the results are presented in Tab. 3. Both taggers are proficient in assigning high-quality tags for the given data in a zero-shot setting. A closer comparison of the results of the two paradigms shows that the selective tagger performs slightly worse than the generative tagger, mainly due to the different stages of tag matching. The selective tagger takes all textual clues as input to perform coarse tag matching before the LLM, which may lead to too many input features and exclude some correct fine-grained tags from the candidate set. However, the generative tagger can effectively retain some detail tags, so it is bound to be more in line with the ground truth.

3.4 Case Study

To provide a more intuitive understanding of TagGPT's performance in zero-shot tagging, this section will present additional case studies to illustrate qualitative results. The specific outcomes are displayed in Tab. 4.

Upon comparing the tagging results from various sources in Tab. 4, it becomes apparent that human tagger often assigns relatively vague and limited tag semantics when dealing with multimodal data. This is usually because micro video creators cannot quickly and accurately find the semantic tags

they need in the massive semantic tags when uploading their works. Moreover, since the popularity of tags directly affects the recommendation algorithm in micro video platforms, creators usually tend to choose popular but ambiguous tags to boost the exposure of their works and reach a broader audience.

Key Frames	Human Tagger	Selective Tagger	Generative Tagger	ACSV	Hanlp
	iPhone小技巧, iCloud空间不足, 苹果手机	存储空间管理, 苹果设备管理, iCloud空间不足, 高新数码产品, iPhone常见问题解决	iCloud空间管理, 苹果手机技巧, 苹果设备管理, iCloud存储空间, iCloud备份与恢复	文本, 手机, 屏幕截图, 小工具, 多媒体	云备份, 储存空间, 实用技巧
	iPhone tips, insufficient iCloud space, iPhone	storage space management, Apple device management, iCloud insufficient space, high-tech digital products, iPhone common problem solving	iCloud space management, Apple mobile phone skills, Apple device management, iCloud storage space, iCloud backup and recovery	text, mobile, screenshot, gadgets, multimedia	cloud backup, storage space, practical skills
	装修, 设计, 公寓装修	家居DIY, 多功能房间设计, 单身公寓设计, 小户型设计, 实用家居	迷你公寓设计, 小户型设计, 装修技巧, 创意装修, 居家便捷设计	墙壁, 室内, 门, 瓷砖, 室内设计	鞋柜, 书柜, 创意设计, 公寓
	decoration, design, apartment decoration	home DIY, multifunctional room design, single apartment design, small apartment design, practical home furnishing	mini apartment design, small apartment design, decoration skills, creative decoration, convenient home design	wall, interior, door, tile, interior design	shoe cabinet, bookcase, creative design, apartment
	比亚迪, 新能源汽车, 唐dmi冠军版	比亚迪唐, SUV汽车, 汽车销量分析, 豪华七座SUV, 比亚迪汽车	唐DM-i冠军版, 新能源汽车, 比亚迪唐DM-i, 豪华SUV, 豪华七座SUV	车辆, 陆地车辆, 文本, 衣服, 汽车设计	老款, 冠军版, 快充, 尊贵, 比亚迪
	BYD new energy vehicle,Tang dmi champion edition	BYD Tang, SUV car, car sales analysis, luxury seven-seat SUV, BYD car	Tang DM-i champion version, new energy vehicles, BYD Tang DM-i, luxury SUV, luxury seven-seat SUV	vehicle, land vehicle, text, clothes, car design	old model, champion version, fast charge, distinguished, BYD
	快来钓鱼, 快手野钓月, 哈利路亚, 快手钓鱼	休闲钓鱼, 钓鱼心得, 钓鱼乐趣, 钓鱼爱好者, 生活秀	钓鱼乐趣, 钓鱼放松, 娱乐休闲, 渔具装备, 钓技攻略	文本, 运动, 水, 海报, 钓鱼	钓鱼, 渔获, 算账, 快手
	come and fish, Kuaishou wild fishing month, Hallelujah, Kuaishou fishing	leisure fishing, fishing experience, fishing fun, fishing enthusiasts, life show	fishing fun, fishing relaxation, entertainment and leisure, fishing gear and equipment, fishing skills raiders	text, sports, water, poster, fishing	fishing, fishery harvesting, accounting, kuaishou
	花卷, 花样美食制作, 面食	花卷制作, 自制面食技巧, 美食掌故, 家庭蒸菜, 花样面食	花卷制作, 食用小方法, 精选主食做法, 葱油花卷制作技巧, 早餐好吃的面食	食品, 蔬菜	蒸葱, 拌均匀, 花卷, 面食
	steamed twisted roll, fancy food production, cooked wheaten food	steamed twisted roll production, homemade pasta skills, gourmet stories, home steamed dishes, pattern pasta	steamed twisted roll production, eating tips, selected staple food practices, scallion oil steamed twisted roll making skills, delicious pasta for breakfast	food, vegetables	Steamed shallots, stir evenly, steamed twisted roll, cooked wheaten food

Table 4: Qualitative results of TagGPT. ACSV and Hanlp are https://portal.vision.cognitive.azure.com/demo/generic-image-tagging and https://www.hanlp.com/demonstrate.html.

A further comparison of the two different paradigm tagging models in TagGPT reveals that both the selective tagger and the generative tagger can produce better tagging results than the human tagger. The selective tagger selectively assigns tags based on a subset of the roughly matched tags, leading to some degree of semantic coarseness while still being able to match the semantics of the given data. On the other hand, the generative tagger understands the given multimodal data and assigns tags in a more precise and accurate manner, fitting the content details more seamlessly. The selective tagger versus generative tagger comparison is essentially a trade-off between universality and precision in the tag assignment process. Overemphasizing universality can result in tags that are indistinguishable and lack analytical value, while overemphasizing precision can lead to the risk of over-interpretation, making it difficult to correctly associate relevant content.

4 Limitations

It has been verified that TagGPT is capable of automatically constructing a tagging system from large-scale multimodal data and providing tags for given data in a zero-shot setting. However, due to its design and technology, it is still subject to certain limitations:

Limitation of LLM Intuitively, the understanding and generation abilities of the LLM are directly related to the performance of TagGPT. Therefore, when the performance of the LLM decreases, TagGPT's performance will also be affected accordingly. Hence, it is recommended to utilize an LLM with robust performance or a language model that has undergone prompt tuning for optimal results.

Limitation of input length When the provided multimodal data contains excessive information, it can result in an excessively long input for the LLM, leading most LLMs to perform truncation operations that render the data unusable. This phenomenon renders TagGPT ineffective in situations involving lengthy videos or texts.

Limitation of security and privacy As a lot of LLMs and unsupervised representation acquisition methods are derived from the API interface within the network, uploading a large amount of user and other data without caution may lead to significant security and privacy concerns regarding data leakage.

5 Conclusion

We propose TagGPT, a concise and low-coupling framework that provides a complete set of tools for tag system construction and tagging models. This framework enables users to easily obtain content tags of large-scale multimodal data and perform zero-shot tagging, which can serve various downstream applications such as search and recommendation. To build a high-quality tag system based on understanding multimodal content, we introduce an LLM and an unsupervised text representation model. We also propose two different zero-shot tagging paradigms, both of which yield high-quality tagging results. In the future, we aim to further optimize each sub-module in TagGPT to improve its work efficiency and accuracy.

Acknowledgements

The project of TagGPT is still under development with the aim of an industrial-strength zero-shot tagger. TagGPT is a collaborative work of ARC-QQ Joint Lab at Tencent PCG, with the help of Zekun Wang, Yunxuan Zhang, Kun Yi, Shupeng Su, and Shansong Liu.

A Instruction template

Dataset	Instruction template
Kuaishou	你是一个视频的兴趣标签生成机器人，根据输入的视频标题、类别、ocr、asr推理出合理的"兴趣标签"，以多个多于两字的标签形式进行表达，以顿号隔开。例如，给定一个视频，它的"标题"为"全屋嵌入式低音音响，主要是这个投影仪真的是爱了"，"类别"为"房产家居"，"ocr"为"42平,一室一厅小户型"，"asr"为"看，远方灯火闪亮着光。你一人低头在路上。这城市越大，越让人心慌多向往，多漫长。祝一路行李太多伤。把最初笑容都淡忘。时光让我们变得脆弱，却坚强，让我在爱青青对你唱。我多想能多陪你唱。把什么生的风景对你讲。"，兴趣标签生成机器人推断出合理的"兴趣标签"为"小户型装修、一室一厅装修、装修效果图"。那么，给定一个新的视频，它的"标题"为"{title}"，"类别"为"{category}"，"ocr"为"{ocr}"，"asr"为"{asr}"，请推断出该视频的"兴趣标签"：
Food.com	You are a tag generation bot for a recipe video, which infers reasonable recipe tags based on the input video "dish name", "description" and "asr", and expresses it in the form of tags with more than two words, separated by commas. For example, given a recipe video whose "dish name" is "Grilled Garlic Cheese Grits", "description" is "We love grits, this is another good way to serve them. A great alternative to a baked potato when served with grilled steak or chicken. I believe this recipe could be made with instant grits. The 2 1/2 hours for refrigeration is not included in the time. The recipe comes from Taste of Homes Light and Tasty.", and "asr" is "In a saucepan, bring water to a boil; slowly add grits and salt, stirring constantly; Reduce heat: simmer, uncovered, for 40-45 minutes or until thickened, stirring occasionally. Add cheese and garlic; stir until cheese is melted, Spray a 9-inch baking dish with nonstick cooking spray; Cover and refrigerate for 2 to 2 1/2 hours or until firm. Before starting the grill, coat the grill rack with nonstick cooking spray; Cut the grits into 3-inch squares; Brush both sides with olive oil. Grill, covered, over medium heat for 4 to 6 minutes on each side or until lightly browned.", the tag generation bot infers a plausible recipe tag as "time-to-make, course, main-ingredient, preparation, occasion, side-dishes, eggs-dairy, refrigerator, diabetic, vegetarian, grains, cheese, stove-top, dietary, low-cholesterol, low-calorie, comfort-food, low-carb, low-in-something, pasta-rice-and-grains, brunch, taste-mood, equipment, presentation, served-hot, 4-hours-or-less". Then, given a new video whose "dish name" is "{name}", "description" is "{caption}", and "asr" is "{asr}", please infer that the videos recipe tag is:

Table 5: Instruction template for tag generation in TagGPT.

In this section, we show the instruction template used for different languages and different datasets as shown in Table 5.

References

[1] Sajad Ahmadian, Milad Ahmadian, and Mahdi Jalili. A deep learning based trust-and tag-aware recommender system. *Neurocomputing*, 488:557–571, 2022.

[2] Yali Amit, Pedro Felzenszwalb, and Ross Girshick. Object detection. *Computer Vision: A Reference Guide*, pages 1–9, 2020.

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz

Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020.

[4] Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, Soumya K Ghosh, Arindam Chaudhuri, Krupa Mandaviya, Pratixa Badelia, and Soumya K Ghosh. *Optical character recognition systems*. Springer, 2017.

[5] Aakanksha Chowdhery, Sharan Narang, and et al. Palm: Scaling language modeling with pathways, 2022.

[6] Maarten Clements, Arjen P de Vries, and Marcel JT Reinders. The influence of personalization on tag query length in social media search. *Information Processing & Management*, 46(4):403–412, 2010.

[7] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. *arXiv preprint arXiv:2104.08821*, 2021.

[8] Manish Gupta, Rui Li, Zhijun Yin, and Jiawei Han. Survey on social tagging techniques. *ACM Sigkdd Explorations Newsletter*, 12(1):58–72, 2010.

[9] Fouzia Jabeen, Shah Khusro, Amna Majid, and Azhar Rauf. Semantics discovery in social tagging systems: A review. *Multimedia Tools and Applications*, 75:573–605, 2016.

[10] Yu Kong and Yun Fu. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5):1366–1401, 2022.

[11] Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. Rethinking the role of demonstrations: What makes in-context learning work? *arXiv preprint arXiv:2202.12837*, 2022.

[12] Liqiang Nie, Leigang Qu, Dai Meng, Min Zhang, Qi Tian, and Alberto Del Bimbo. Search-oriented micro-video captioning. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 3234–3243, 2022.

[13] OpenAI. Gpt-4 technical report, 2023.

[14] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback, 2022.

[15] Priyanka Panchal and Dinesh J Prajapati. The social hashtag recommendation for image and video using deep learning approach. In *Sentiment Analysis and Deep Learning: Proceedings of ICSADL 2022*, pages 241–261. Springer, 2023.

[16] Filip Radlinski, Craig Boutilier, Deepak Ramachandran, and Ivan Vendrov. Subjective attributes in conversational recommendation systems: challenges and opportunities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pages 12287–12293, 2022.

[17] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

[18] Zhenhailong Wang, Manling Li, Ruochen Xu, Luowei Zhou, Jie Lei, Xudong Lin, Shuohang Wang, Ziyi Yang, Chenguang Zhu, Derek Hoiem, et al. Language models with image descriptors are strong few-shot video-language learners. *arXiv preprint arXiv:2205.10747*, 2022.

[19] BigScience Workshop, :, Teven Le Scao, Angela Fan, and et al. Bloom: A 176b-parameter open-access multilingual language model, 2023.

[20] Zhichen Xu, Yun Fu, Jianchang Mao, and Difu Su. Towards the semantic web: Collaborative tag suggestions. In *Collaborative web tagging workshop at WWW2006, Edinburgh, Scotland*, 2006.

[21] Dong Yu and Li Deng. *Automatic speech recognition*, volume 1. Springer, 2016.

[22] Andy Zeng, Maria Attarian, Brian Ichter, Krzysztof Choromanski, Adrian Wong, Stefan Welker, Federico Tombari, Aveek Purohit, Michael Ryoo, Vikas Sindhwani, Johnny Lee, Vincent Vanhoucke, and Pete Florence. Socratic models: Composing zero-shot multimodal reasoning with language, 2022.

[23] Yang Zhang, Fuli Feng, Xiangnan He, Tianxin Wei, Chonggang Song, Guohui Ling, and Yongdong Zhang. Causal intervention for leveraging popularity bias in recommendation. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 11–20, 2021.