KnowBLIP: Large Vision-Language Models meet Interleaved World Knowledge

Anonymous Author(s)
Affiliation
Address
email

Abstract

Currently, Large Visual Language Models (LVLMs) have integrated various tasks, and some "any2any" models have emerged, enabling support for diverse modalities for input and various output formats, thereby expanding the impact of LVLMs. However, we note that existing models remain incomplete and distant from real-world applicability, primarily due to two key factors: partial functionality and lack of connection with world knowledge. Firstly, we stress the importance of a comprehensive LVLM that accommodates various input and output paradigms, which should be interleaved, multi-turn, and multi-lingual. However, there is no multi-modal framework capable of fully encompassing all functionalities. Moreover, textual knowledge often lacks substantial connections with the visible objective world, resulting in visual and language alignment limited to the semantic level, detached from objective facts. This limitation leads to models generated primarily through imagination rather than a grounded understanding of the world. To bridge the gap between LVLMs and the visible real world, we propose a novel multi-modal framework, KnowBLIP, designed to support a comprehensive, world-grounded, "any2any" paradigm. Specifically, we have collected a dataset, IWK-500k, of multi-modal objective knowledge derived from reasoning on multi-modal knowledge graphs, characterized by interleaved, multi-turn, and multi-lingual attributes. Furthermore, we have developed the KnowBLIP, capable of accommodating various input and output paradigms based on IWK-500k dataset. Extensive experiments have demonstrated the effectiveness of our approach, and both the dataset and code will be made publicly available[1].

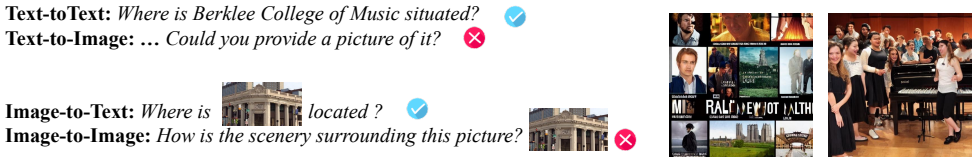
1 Introduction

With the development of the internet and social media, data across various modalities has seen explosive growth. Harnessing this vast amount of data, *unimodal* models have achieved groundbreaking successes in their respective domains, such as the text-oriented ChatGPT[2] and the vision-oriented ViT[1]. However, their lack of exposure to data from other modalities renders them fall short in multi-modal tasks. To address this *multimodal* challenge, **L**arge-scale **V**ision-**L**anguage **M**odel**s** (**LVLMs**) like GPT4-V and LLaVA[2] have emerged. By understanding, integrating, and aligning data from different modalities, LVLMS have demonstrated remarkable performance across numerous multimodal tasks.

Along this line, we revisit the aforementioned pioneer LVLMs: MiniGPT-4[3] and LLaVA[2] (*Visual Question Answering*), Shikra[4] and LiSA[5] (*Grounding and Segmentation*), MiniGPT-5[6] and

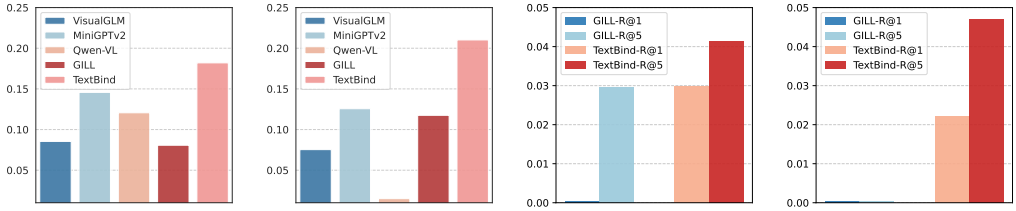
[1] https://drive.google.com/drive/folders/1ZgfApnIhzXg8n3R_kwikix5C0Klv6UP8
[2] https://openai.com/blog/chatgpt

Submitted to 38th Conference on Neural Information Processing Systems (NeurIPS 2024). Do not distribute.



(a) Interleaved Question revolves around World Knowledge. (b) Generated images by GILL.

Figure 1: The four ideal paradigms exist within LVLMs. However, most baselines are unable to generate images. Even models that can generate images, like GILL, struggle to obtain objective results, as demonstrated in the case of "Boston".



(a) Accuracy on T2T. (b) Accuracy on I2T. (c) Recall on T2I. (d) Recall on I2I.

Figure 2: The zero-shot testing of world knowledge answering on exsiting models. The accuracy of knowledge and the similarity between generated images and real images are insufficient.

DreamLLM[7] (*Text-to-Image Generation*). Despite these LVLMs advancing the field of multimodal tasks, it's regrettable that they are **"Non-any2any" multimodal models** (i.e., *not supporting arbitrary input and output modalities*) but only a subset of "any2any" modes (e.g., T2I, I2T) as shown in Table1. Considering that humans communicate in an "any2any" mode in real life, the "any2any" paradigm is essential path toward multimodal general models. In light of this, researchers have focused and proposed several "any2any" models, such as TextBind[8], NExT-GPT[9], to aid LVLMs in their real-world practices. Attracted by the powerful capabilities of existing "any2any" models, we employed these models to accomplish multimodal tasks, but discover a common issue — *the generated content do not align with the objective facts of the real world* — we refer to as **"A gap with world knowledge"**. For example, as shown in Figure1, when addressing the question "Where is the Berklee College of Music located?", the text generated by VisualGLM, the image produced by GILL[10], or the text-image content generated by TextBind[8], are all incorrect. Further, to verify the "A gap with world knowledge" issue indeed exists in current "any2any" baselines, we meticulously conducted four multimodal world knowledge QA tasks. As illustrated in Figure2, *the poor performance* of text generation ($\leq$25%) and image generation ($\leq$5%) tasks *validate that existing "any2any" models rely on imagining rather than understanding world to generate content*. Therefore, we underscore the importance of integrating world knowledge into "any2any" multimodal models, thereby facilitating the generation of content that aligns with objective facts.

Facing the challenges of *"Non-any2any"* and *"A gap with world knowledge"*, in this paper, we develop a comprehensive multimodal benchmark to foster a close connection between "any2any" LVLMs and the real world: 1)-we collected a multimodal objective knowledge dataset **IWK-500**, via multimodal knowledge graph (MMKG) reasoning, as depicted in radar chart 5. Notably, to bring this dataset closly mimic real-world QA scenarios, we added three key attributes: *i)-Interleaved*: both queries and answers contain text and images; *ii)-Multi-turn*: the query-answer pairs support the form of multi-turn dialogues; *iii)-Multilingual*: the query-answer pairs support multiple languages. 2)-we developed an "any2any" LVLM **KnowBLIP** driven by autoregressive text generation and contrastive-based image retrieval tasks, as illustrated in the framework diagram 6.

Our contributions are summarized as follows:

1. We pioneered a comprehensive multimodal benchmark with the creation of the IWK-500 dataset, a **first-of-its-kind** that supports *"any2any", interleaved, multi-turn, multi-lingual, and incorporates world-knowledge aspects*. Leveraging this, we introduced "any2any" LVLMs, KnowBlip, designed for *practical use and aligned with human needs*.
2. Over **six months**, we meticulously prepared prompts and engaged **10 annotators** to compile the IWK-500 dataset. KnowBlip then outperformed **8 leading baselines**, such as GPT-4V

Table 1: Comparision between our collected datasets and other existing datasets. Among them, the "T2I" means Text-to-Image, aka image generation.

Dataset	Type	Multi-Trun	Interleaved	World Knowledge	Multi-lingual	Image volume	Text volume
ShareGPT52K	T2T	✔	✗	✗	✔	-	90K
Alpaca-GPT4[11]	T2T	✗	✗	✗	✔	-	52K
Open Assistant	I2T	✗	✗	✗	✔	-	84K
COCO Caption[12]	I2T	✗	✗	✗	✔	164K	1M
CC12M[13]	I2T	✗	✗	✗	✗	12M	12M
LAION-COCO[14]	I2T	✗	✗	✗	✗	600M	300M
VQAv2[15]	I2T & T2T	✗	✗	✗	✗	265K	1.4M
OK-VQA[16]	T2I & T2T	✗	✗	✗	✗	14K	14K
GQA[17]	T2I & T2T	✗	✗	✗	✗	113K	22M
Visual Genome[18]	T2I & T2T	✗	✗	✗	✗	108K	1.77M
ScienceQA[19]	T2I & T2T	✗	✗	✗	✗	10k	21K
TextVQA[20]	T2I & T2T	✗	✗	✗	✗	28k	45K
LLaVA-instruct-150K[2]	T2I & T2T	✔	✗	✗	✗	82K	150K
MMDialog[21]	T2I & I2T & T2T & I2I	✔	✔	✗	✗	1.53M	1.08M
Ours	T2I & I2T & T2T & I2I	✔	✔	✔	✔	650K	1.2M

and Qwen-VL, on the xxxx task, setting a **new record for state-of-the-art performance** in downstream multimodal tasks.

2 Related Work

2.1 Large Vision-Language Models

Recent progresses of computational resources has greatly facilitated the research into large-scale foundational models incorporated with multi-modal learning [22]. Powered by open sourcing large language models such as LLaMA[23] and Vicuna[24], LVLMs understand and generate diverse content in a more comprehensive way by integrating information from different modalities, such as text, images, and audio. Qwen-vl[25], VisualGLM[26] and MiniGPT-4[3] take a step forward in this field, allowing users to interact with these intelligence with images and texts as prompts. All of them share the same two training phases, i.e., pre-trained feature alignment and instruction fine-tuning[27], to help the model to comprehend the format of instruction input. With the assist of Diffusion model[28], LVLMs are extended to image generation task. GILL[10], TextBind[8] translates hidden representations of text into the embedding space of the visual models, enabling LVLMs to leverage the strong text representations of the LLM for visual outputs. However, all of aforementioned LVLMs suffer from lacking connection with the visible world. Consequently, we mainly conduct the experiments on these models to solve this problem in our paper.

2.2 Multi-modal Knowledge Graph

In this paper, we will utilize a multi-modal knowledge graph (MMKG) to establish connections between text and the physical world[29]. A knowledge graph (KG) provides a structured representation of knowledge, typically in the form of triples consisting of (head, predicate, tail). In this schema, the head and tail denote entities, while the predicate indicates the type of relationship between them. In the context of MMKG, entities can also encompass images and other modalities, for instance (France, hasImage, <IMG>), where <IMG> represents the flag of France. This framework facilitates the seamless linkage of abstract textual concepts to tangible real-world entities, enabling the model to transcend reliance on abstract notions and establish concrete associations with the physical realm.

3 Proposed Method

To facilitate the LVLM in acquiring objective knowledge of the visible real world, we collect a comprehensive dataset by reasoning on MMKG. Moreover, in Section 3.2, we will elaborate on how our framework achieves comprehensive and world-grounded performance.

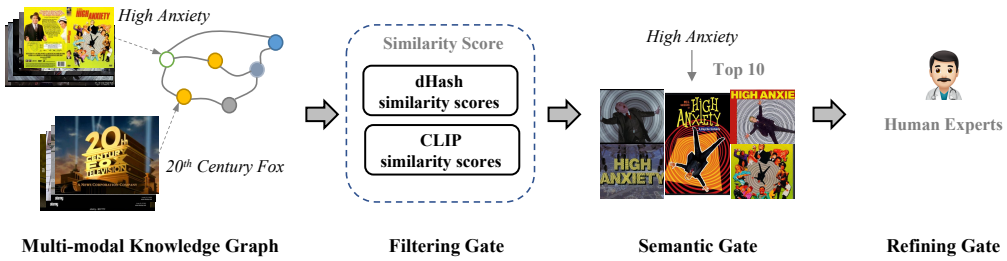


Figure 3: The pipeline for selecting a image from Multi-model Knowledge Graph.

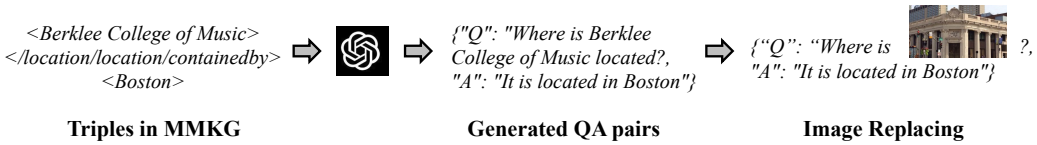


Figure 4: The pipeline for generating interleaved images.

3.1 IWK-300K Collection

We generated a comprehensive dataset of factual knowledge using the multi-modal knowledge graph (MMKG), which is built upon FB15K[29] to create interleaved data in diverse forms, including Text-to-Text (T2T), Image-to-Text (I2T), Text-to-Image (T2I), and Image-to-Image (I2I). In FB15K, knowledge is structured into triples (head entity, relation, tail entity), where each entity has multiple images. To select the most representative image, as depicted in Figure 3, three gates are designed: the filter gate, semantic gate, and refining gate. The filter gate utilizes dHash and CLIP[30] to compute image similarity, thereby retaining images with minimal repetition. Subsequently, the semantic gate selects the most relevant images to the textual entity. Finally, expert selection is employed to choose an image to represent the entity. Following rigorous endeavors, we successfully obtained a dataset of $298,957$ triplets, including $14,505$ entries.

As shown in Figure 4, in order to produce QA pairs suitable for LVLMs training, we employed GPT-4[31] to fashion a natural language description of triples, as well as polished QA pairs derived from the subsequent prompt.

> **Input:** Please generate QA pair for the description "Berklee College of Music is located in Boston" and return them in JSON format. Such as ...
> **Output:** {"Q": "Where is Berklee College of Music located?, "A": "It is located in Boston" } ...

Thereafter, entity concepts are randomly replaced by corresponding images with a 50% probability, forming pairs like (<IMG> of head entity, relation, tail entity).

Furthermore, recognizing the significance of multi-turn conversations in enhancing the few-shot[32] capability of LVLMs, we have curated multi-turn data by consolidating identical entities. Notably, every output produced by GPT-4 will undergo meticulous scrutiny and refinement by 10 human experts to guarantee the precision of knowledge. Ultimately, we acquired a multi-turn, interleaved dataset comprising 650K images and 1.2M text volumes, named IWK-300k.

Statistics We have compiled fundamental attributes of IWK-300k, summarized in Table 2. The dataset comprises $298,957$ rounds of conversation, averaging 3.49 rounds per complete conversation. Additionally, we conducted an analysis of the occurrence of "wh" words in QA pairs to ensure question diversity. Notably, owing to the rich location attributes in MMKGs, "who" questions are predominant, totaling $136,862$ occurrences. From a visual standpoint, the average number of images per

Table 2: Statistics of IWK-300k dataset.

Total datas	298,957
Total entitys	14,505
Avg. #data/entity	20.61
Avg. #image/entity	61.23
Avg. #turns in conversations	3.49
Total number of QA pairs	
"What"	124,111
"When"	9,271
"Which"	5,669
"Who"	136,862
"Where"	21,948
"Is/Are/Did"	1,096

(a) The distribution of topics.

(b) The distribution of sentence length.

Figure 5: The statistics in our IWK-300k dataset.

entity is 61.23, with each entity featuring in an average of 20.61 questions. Lastly, we compared the performance of GPT-4 before and after refinement and enhancement, illustrated in Figure 5. The length of questions per round increased from an average of 10 words to over 18 words. More details are disclosed in the Appendix A.

Diversity To evaluate sentence diversity within the IWK-300k datasets, we perform a topic frequency analysis of world knowledge present in the IWK-300k corpus. As illustrated in Figure 5, we initially quantify the total occurrence of each topic among entities. Among them, there are 63 distinct topic types in total. Through human classification and summarization, these topics are categorized into 12 areas, encompassing nearly comprehensive and diverse real-world domains.

Formally, we define this dataset constructed as D=$\{(X_u^1, Y_m^1), ..., (X_u^n, Y_m^n)\}$, where the X_u^i represent the i-th instruction from user while the Y_m^i represent the i-th answer from model. Meanwhile, a interleaved input X_u^i is defined as X_u^i= $\left(x_t^1, x_t^2, x_i^1, ..., x_t^n, x_i^k\right)$, where x_t^* represents the text token of multi-turn conversations , x_i^* represents the visual part interleaving in the conversations. The n denotes the total number of text tokens, while the k denotes the total number of images in conversations.

3.2 KnowBLIP Framework

The overall training framework is depicted with an illustration in Figure 6. In line with most LVLMs, our framework employs ViT-L/14 from BLIP-2[33] as the visual encoder, a pretrained Q-Former as the visual adapter, and LLaMA-2-7B as the LLM. During training, the adapter undergoes tuning[34] while the ViT and LLM remain frozen.

On the training function, we extended the autoregressive loss $\mathcal{L}_{AR}$ of the language model. Formally, given a question in a single turn conversation X_u^i= $\left(x_t^1, x_t^2, x_i^1, ..., x_t^n, x_i^k\right)$, the model expects to output the answer Y_m^i= $\left(y_t^1, y_t^2, ..., y_t^k, y_t^n\right)$.

$$\mathcal{L}_{\mathrm{AR}} = -\log p(\mathrm{y_t^k} | X_u^i, y_t^j, j < k). \tag{1}$$

To bolster the model's capability in generating objective knowledge, particularly in visual generation, we define the image generation as a retrieval task. This necessitates the LVLM to not only output textual knowledge (i.e. Boston), but also to "generate" Boston images via retrieval process. Specifically, we introduce two special tokens: the knowledge token and the retrieval token, represented as <KNO> and <RET>, for knowledge generation and image retrieval, respectively.

For knowledge generation task, we use <KNO> token as an anchor to mark the world knowledge entity and compute a knowledge loss $\mathcal{L}_{KNO}$. Specifically, a question and answer in a single turn conversation are represented as X_u^i= $\left(x_t^1, x_t^2, x_i^1, ..., x_t^n, x_i^k\right)$ and Y_m^i= $\left(y_t^1, y_t^2, ..., y_{kno}, y_t^n\right)$, where y_{kno} denotes the marked world knowledge entity <KNO>. This <KNO> token will be computed loss with the correctly world knowledge entity:

$$\mathcal{L}_{\mathrm{KNO}} = -\log p(\mathrm{y_{kno}} | X_u^i, y_t^j, j < k), \tag{2}$$

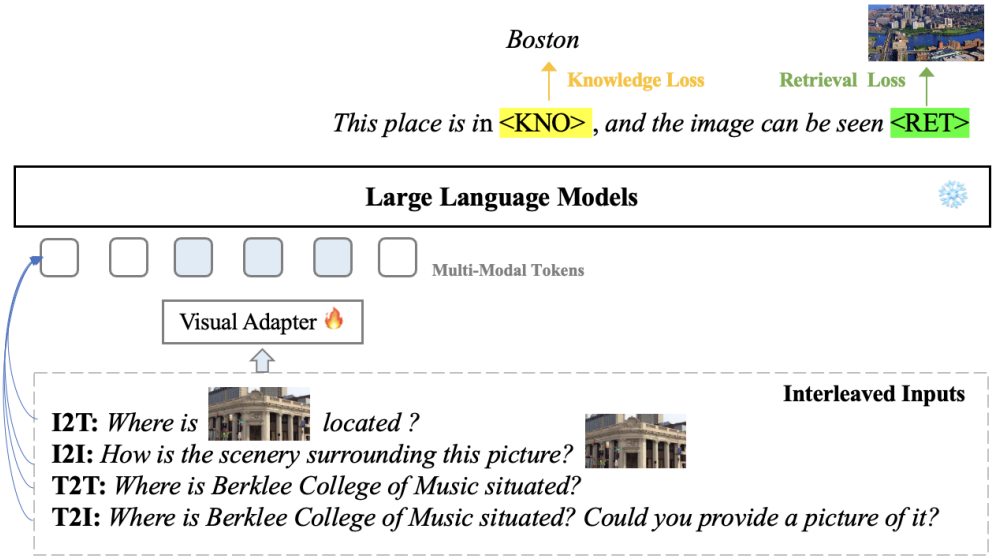


Figure 6: The overview of KnowBLIP for interleaved inputs, which consists of three losses: regression loss $\mathcal{L}_{AR}$, knowledge loss $\mathcal{L}_{KNO}$, and retrieval loss $\mathcal{L}_{RET}$.

Through this loss, the model will be guided to focus on generating correct knowledge.

For image retrieval task, we introduce a specific token <RET>, which serves to "generate" objective images. In the training phase, this token is affixed to the textual context in order to tag a objective image. Correspondingly, during the inference stage, if the model generates a <RET> token, our system proceeds with the retrieval process. The computation of the retrieval loss $\mathcal{L}_{\text{RET}}$, hinges on the ability of the representation of the <RET> token to retrieve images. Similar to CLIP, we employ contrastive loss for this task. Specifically, the embeddings of all potential images are acquired via ViT-L/14 and then mapped onto a unified space through the projection matrix, $W_{img} \in \mathbb{R}^{768\times p}$, where p is the linear dimension of the <RET> vector. Subsequently, the image feature E^i_{img} and the token feature E^i_{ret} are used to calculate contrastive loss[35],

$$\mathcal{L}_{\text{RET}} = -\log \frac{exp(sim(E_{ret}, E^i_{img})/\tau)}{\sum_{j=1}^{n} exp(sim(E_{ret}, E^j_{img}/\tau)}, \tag{3}$$

$$sim(E_{ret}, E^i_{img}) = \frac{(E_{ret})(E^i_{img})^T}{||E_{ret}|| \quad ||E^i_{img}||}, \tag{4}$$

where i represent the matching image and j denotes negative images that is randomly sampled, τ is the temperature coefficient.

Finally, our training function $\mathcal{L}$ consists of three parts, namely regression loss $\mathcal{L}_{AR}$, knowledge loss $\mathcal{L}_{KNO}$, and retrieval loss $\mathcal{L}_{RET}$,

$$\mathcal{L} = \mathcal{L}_{AR} + \lambda_1 \mathcal{L}_{KNO} + \lambda_2 \mathcal{L}_{RET}, \tag{5}$$

where λ_1 and λ_2 are balance parameters.

4 Experiments

Our KnowBLIP is a LVLM designed to incorporate world knowledge, and thus our experiments primarily center on assessing its capacity to produce world knowledge images and texts. In order to compare the effectiveness of our model with existing baselines in terms of world knowledge, we carried out experiments on four subsets of IWK-300k (e.g., Text2Image).

4.1 Datasets

Our datasets are split into 3 non-overlapping subsets, where 0.6, 0.2 and 0.2 are used for training, validation and testing. Specifically, for one triplet that has four formats in different formats of

Table 3: Zero-shot testing of baselines on IWK-300k. Among these, "–" indicates not applicable, as the model is unable to execute the relevant task.

	Text2Text	Image2Text	Text2Image R@1	Text2Image R@5	Image2Image R@1	Image2Image R@5
VisualGLM[26]	8.54%	7.53%	–	–	–	–
Qwen-VL[25]	14.57%	12.56%	–	–	–	–
MiniGPTv2[37]	12.06%	1.50%	–	–	–	–
GILL[10]	8.07%	11.75%	0.00%	2.95%	–	–
TextBind[8]	18.20%	21.02%	2.98%	4.13%	2.21%	4.70%
LLaVA-1.5-7B[38]	6.69%	10.20%	–	–	–	–
CogVLM-chat-v1.1[39]	14.45%	17.41%	–	–	–	–
GPT-4V	61.53%	55.88%	–	–	–	–
Gemini-1.5-pro[40]	56.41%	41.53%	–	–	–	–

Table 4: Performance comparison of baselines after training on IWK-300k. Among these, "–" indicates not applicable, as the model is unable to execute the relevant task. **Bold** represents optimal performance, while underline represents suboptimal.

	Text2Text	Image2Text	Text2Image R@1	Text2Image R@5	Image2Image R@1	Image2Image R@5
VisualGLM[26]	50.72%	22.50%	–	–	–	–
Qwen-VL[25]	51.69 %	48.74%	–	–	–	–
MiniGPTv2[37]	40.62%	30.57%	–	–	–	–
GILL[10]	47.38%	70.03%	6.79%	24.30%	–	–
TextBind[8]	59.38%	64.29%	5.66%	6.30%	3.72%	6.79%
KnowBLIP (Ours)	**66.73%**	**70.60%**	**29.71%**	**78.24%**	**17.47%**	**64.38%**

question-answer pair, we would split its different formats into training, validation and testing sets, which ensures the world knowledge of this triplet having trained.

4.2 Evaluation Metrics

The evaluation is carried out in two aspects: the accuracy of the generated knowledge and the quality of the generated images. These metrics are calculated using <KNO> token and <RET> token. In our framework, the <RET> token will be utilized to retrieve images, while in the baselines, the embedding of the final generated image will be used for retrieval. For knowledge accuracy, the accuracy will be employed for T2T and I2T tasks to ascertain whether the response aligns with the world knowledge entity. Regarding image, recall@1 and recall@5 are utilized to assess whether the images obtained from I2I and T2I tasks are sufficiently objective.

4.3 Implementation Details

In the retrieve process, feature linear mapping W_{RET} and linear mapping W_{img} are 768-dimensional vectors. We use the Adam optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$ for all gradient-based methods, where the mini-batch size and the learning rate were searched and are set up to 16 and 10^{-4}, respectively. We iteratively trained 20K steps on a server equipped with 4 A100 GPUs[36].

4.4 Overall Performance Comparison

In this subsection, we begin by conducting robustness experiments (i.e. zero-shot testing) on the baselines, followed by an overall comparison of the trained baselines.

During zero-shot testing, we evaluate a wide range of models, and metrics are not calculated if the corresponding subtasks cannot be executed. Notably, in GILL, we use the <IMG> token for retrieval, while in TextBind[8], the images generated by stable diffusion are utilized. As depicted in Table 3, the following observations are made: 1) Virtually no model inherently supports all four tasks, underscoring the value of developing a comprehensive LVLM. 2) The majority of models exhibit limited knowledge abilities, indicating that baselines often produce outputs without comprehensive world knowledge.

Table 5: Ablation analysis on different subtasks. Among these, "–" indicates not applicable, as the model is unable to execute the relevant task. **Bold** represents optimal performance, while underline represents suboptimal.

	Text2Text	Image2Text	Text2Image R@1	Text2Image R@5	Image2Image R@1	Image2Image R@5
VisualGLM[26]	56.00%	15.46%	–	–	–	–
Qwen-VL[25]	74.19%	29.47%	–	–	–	–
MiniGPTv2[37]	80.99%	17.51%	–	–	–	–
GILL[10]	41.61%	61.12%	39.77%	71.92%	–	–
TextBind[8]	91.00%	64.16%	6.86%	9.97%	3.88%	7.89%
KnowBLIP (Ours)	**91.29%**	**71.32%**	**25.13%**	**72.54%**	**24.67%**	**70.88%**

In overall comparison, we selected a diverse range of state-of-the-art models. These encompass three baselines that are proficient in text generation, Qwen-VL[25], VisualGLM[26], MiniGPTv2[37]; two models for image generation: GILL, TextBind[8]. Specifically, for text generation baselines, we only conduct experiments on T2T and I2T tasks. For Image generation frameworks, we train TextBind[8] within the origin code and retrieve images via the similarity of images generated by stable diffusion.

The experimental results are presented in Table 4, and the following observations can be made:

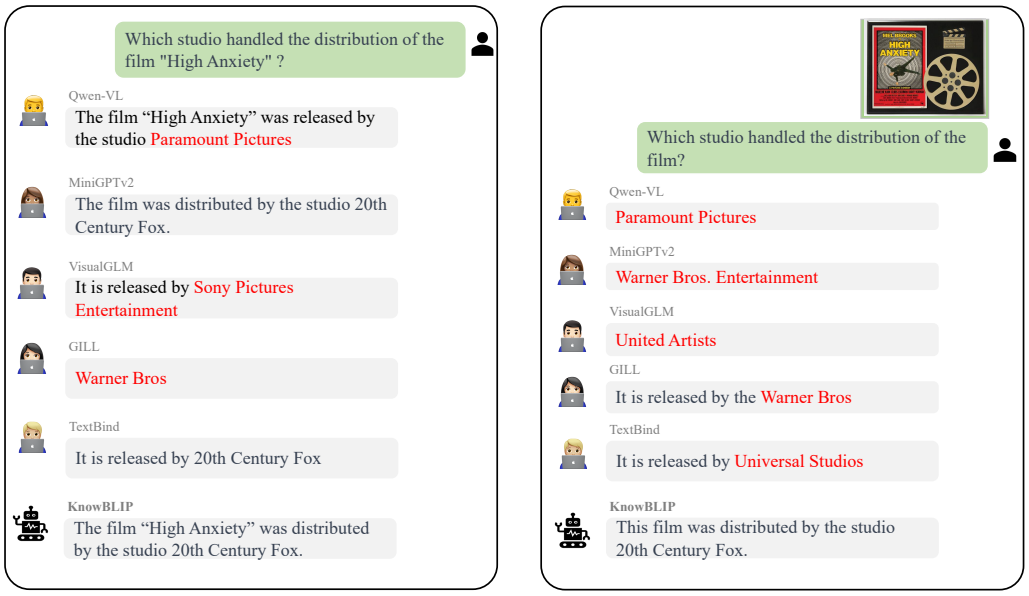
- When training IWK-300k on each model, both models demonstrated a noticeable improvement compared to the zero-shot setting. This improvement suggests that our dataset effectively incorporates world knowledge into this models.
- In the text generation task, all models showed a significant improvement, indicating that textual generation tasks are relatively easy for existing LVLMs. However, it was also observed that the I2T task lagged behind the T2T task, highlighting the challenge of utilizing LVLMs to understand images of world knowledge entities.
- In the image generation task, the model TextBind performed poorly in both the zero-shot and training settings. In the TextBind experiment setting, it encodes the generated text between two special tokens using the text encoder of the SD model to generate an image. However, most of the generated texts are special phrases in IWK-300k, such as the names of celebrities or cities. The images generated by SD still appear to be disconnected from the world knowledge entity, which accounts for the poor performance of TextBind.
- Our **KnowBLIP** exhibited substantial improvement compared to TextBind in all three tasks, especially T2I and I2I. Through the retrieval method, we are able to provide real-world images instead of purely imaginary ones as much as possible.

4.5 Ablation Study

In order to thoroughly investigate our framework and dataset, we conducted ablation studies by training four subsets individually in all models.

The results of the ablation studies are presented in Table 5, and the following observations were made:

- Training individually on each subset task led to a certain degree of improvement in both models compared to the zero-shot setting, indicating that our dataset is capable of injecting world knowledge into the models for each individual task.
- A comparison with the results of mixed training tasks revealed a slight increase in the performance of the I2T task during individual training. Conversely, the T2T task showed a minor decrease in individual training. These observations suggest that mixed training enhances the ability to comprehend input images and better aligns world knowledge entities with the input images.
- However, this setting also has a negative impact on how well models master world knowledge through text-only. Similarly, similar performance differences were observed between I2I and T2I tasks.



(a) Text2Text task. (b) Image2Text task.

Figure 7: The visualization of several models in text generation tasks.

4.6 Visualization

To thoroughly examine our framework's processes during execution, we have employed visualizations to elucidate, as shown in Figure 7 and Figure 8. 1) For the text generation task, all models have the capability to understand the content of the text or image and perform the basic task of question and answering. This is evident in their ability to provide appropriate responses when asked about a film distribution company, such as “Warner Bros” or “Sony Pictures Entertainment”. However, their performance is unsatisfactory when faced with world knowledge in the <KNO> slot that requires a response. This indicates that while all models can easily master the correct response format, they struggle with the correct world knowledge entity slot. Similarly, GILL and TextBind make mistakes in image generation tasks, leading to a chain of errors in the retrieved or generated images. 2) In the image generation task, we observed that the world knowledge of the <KNO> slot is correct in both TextBind and KnowBLIP in the T2I task. However, TextBind generates an image that matches the literal meaning (such as a strange **fox**, possibly caused by the modifiers of **20th Century**). More examples are available in the appendix.

5 Conclusions

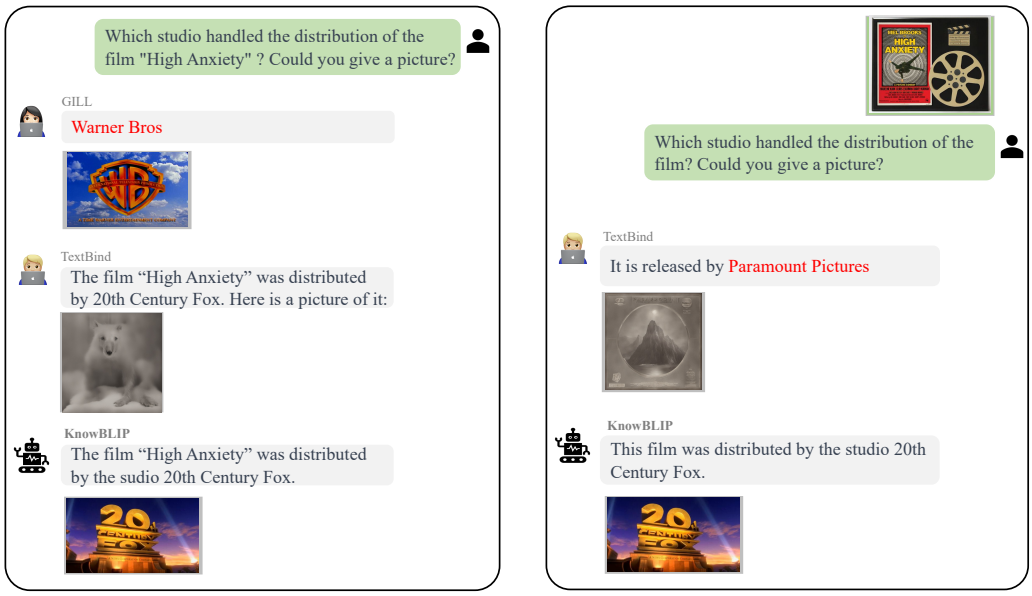
In this study, we aim to explore a collection of datasets on world knowledge in order to enhance the contextual understanding of LVLMs. Our proposed approach, KnowBLIP, facilitates a comprehensive, world-based “any2any” paradigm that bridges the gap between LVLMs and real-world applications. Through zero-shot testing and comprehensive comparative analysis, we have demonstrated the effectiveness of our method.

Moving forward, we plan to expand the scope of this research by increasing the size of the dataset and enhancing the parameter size. We believe that augmenting the volume of data and parameters will enable the model to better align with real-world contexts.

References

[1] Dosovitskiy, A., L. Beyer, A. Kolesnikov, et al. An image is worth 16x16 words: Transformers for image recognition at scale, 2021.

[2] Liu, H., C. Li, Q. Wu, et al. Visual instruction tuning, 2023.



(a) Text2Image task.

(b) Image2Image task.

Figure 8: The visualization of several models in image generation tasks.

[3] Zhu, D., J. Chen, X. Shen, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023.

[4] Chen, K., Z. Zhang, W. Zeng, et al. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023.

[5] Lai, X., Z. Tian, Y. Chen, et al. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*, 2023.

[6] Zheng, K., X. He, X. E. Wang. Minigpt-5: Interleaved vision-and-language generation via generative vokens, 2023.

[7] Dong, R., C. Han, Y. Peng, et al. DreamLLM: Synergistic multimodal comprehension and creation. In *The Twelfth International Conference on Learning Representations*. 2024.

[8] Li, H., S. Li, D. Cai, et al. Textbind: Multi-turn interleaved multimodal instruction-following in the wild, 2024.

[9] Wu, S., H. Fei, L. Qu, et al. Next-gpt: Any-to-any multimodal llm, 2023.

[10] Koh, J. Y., D. Fried, R. Salakhutdinov. Generating images with multimodal language models. *NeurIPS*, 2023.

[11] Peng, B., C. Li, P. He, et al. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*, 2023.

[12] Chen, X., H. Fang, T.-Y. Lin, et al. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[13] Changpinyo, S., P. Sharma, N. Ding, et al. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts, 2021.

[14] Schuhmann, C., R. Vencu, R. Beaumont, et al. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs, 2021.

[15] Goyal, Y., T. Khot, D. Summers-Stay, et al. Making the V in VQA matter: Elevating the role of image understanding in Visual Question Answering. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2017.

[16] Marino, K., M. Rastegari, A. Farhadi, et al. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*. 2019.

[17] Hudson, D. A., C. D. Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709. 2019.

[18] Krishna, R., Y. Zhu, O. Groth, et al. Visual genome: Connecting language and vision using crowdsourced dense image annotations. 2016.

[19] Lu, P., S. Mishra, T. Xia, et al. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*. 2022.

[20] Singh, A., V. Natarajan, M. Shah, et al. Towards vqa models that can read. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8317–8326. 2019.

[21] Feng, J., Q. Sun, C. Xu, et al. MMDialog: A large-scale multi-turn dialogue dataset towards multi-modal open-domain conversation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7348–7363. Association for Computational Linguistics, Toronto, Canada, 2023.

[22] Zeng, Y., D. Cao, X. Wei, et al. Multi-modal relational graph for cross-modal video moment retrieval. In *Proceedings of the CVPR*, pages 2215–2224. IEEE, 2021.

[23] Touvron, H., T. Lavril, G. Izacard, et al. Llama: Open and efficient foundation language models, 2023.

[24] Chiang, W.-L., Z. Li, Z. Lin, et al. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality, 2023.

[25] Bai, J., S. Bai, S. Yang, et al. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

[26] Du, Z., Y. Qian, X. Liu, et al. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 320–335. 2022.

[27] Ouyang, L., J. Wu, X. Jiang, et al. Training language models to follow instructions with human feedback, 2022.

[28] Rombach, R., A. Blattmann, D. Lorenz, et al. High-resolution image synthesis with latent diffusion models, 2022.

[29] Bordes, A., N. Usunier, A. Garcia-Durán, et al. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems (NIPS 26)*. 2013.

[30] Radford, A., J. W. Kim, C. Hallacy, et al. Learning transferable visual models from natural language supervision, 2021.

[31] OpenAI, J. Achiam, S. Adler, et al. Gpt-4 technical report, 2024.

[32] Sanh, V., A. Webson, C. Raffel, et al. Multitask prompted training enables zero-shot task generalization, 2022.

[33] Li, J., D. Li, S. Savarese, et al. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023.

[34] Wei, J., M. Bosma, V. Y. Zhao, et al. Finetuned language models are zero-shot learners, 2022.

[35] He, K., H. Fan, Y. Wu, et al. Momentum contrast for unsupervised visual representation learning, 2020.

[36] Rasley, J., S. Rajbhandari, O. Ruwase, et al. Deepspeed: System optimizations enable training deep learning models with over 100 billion parameters. In *KDD '20: The 26th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*. 2020.

[37] Chen, J., D. Zhu, X. Shen, et al. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *arXiv preprint arXiv:2310.09478*, 2023.

[38] Liu, H., C. Li, Y. Li, et al. Improved baselines with visual instruction tuning, 2024.

[39] Wang, W., Q. Lv, W. Yu, et al. Cogvlm: Visual expert for pretrained language models, 2024.

[40] Team, G., M. Reid, N. Savinov, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024.

[41] Wang, Y., Y. Kordi, S. Mishra, et al. Self-instruct: Aligning language models with self-generated instructions, 2023.

[42] Sun, Q., Y. Wang, C. Xu, et al. Multimodal dialogue response generation, 2022.

[43] Oñoro-Rubio, D., M. Niepert, A. García-Durán, et al. Answering visual-relational queries in web-extracted knowledge graphs. In *AKBC*. 2019.

A Details of IWK-300K Collection

In this section, we will provide a comprehensive overview of dataset collection.

A.1 Prompt for Data Construction

To construct a high-quality dataset and reduce labor costs, we divide the dataset construction pipeline into three subtasks and create detailed prompt templates for semi-automatic data production. The three parts include 1) transforming triplets into natural sentences, 2) writing diverse questions, and 3) constructing multi-turn conversations.

A.1.1 Step 1 - Transforming Triplets into Natural Sentences

We begin by extracting the original data from a multi-modal knowledge graph in the format of **"<head_id, relation_chain, tail_id>"** (e.g. **"</m/027rn, /location/country/form_of_government, /m/06cx9>"**). In this format, **"/m/027rn"** and **/m/06cx9"** represent the IDs of the head entity and tail entity respectively. The **"/location/country/form_of_government"** signifies the relation chain between the head entity and the tail entity. By post-processing the correspondence relationship of the ID-entity, we can obtain plaintext in the format of **"<Dominican Republic, /location/country/-form_of_government, Republic>"**. Due to creating directly from triples to QA pairs, we firstly transforming triplets into natural sentences. Specifically, we randomly select 10 distinct triples and translate into natural sentences by human experts. Then these triples are treated as few shot templates and the following prompt is given to GPT-4 to produce natural sentences[41].

Input:
Task definition: there is triple format information from a knowledge graph. You are required to transform it in natural sentences.
Templates:
Triple 1:< Boston College> < /location/location/containedby > < United States of America > Answer 1:
1. Boston College is located within the United States of America.
2. Boston College is contained by the United States of America.
. . . . . .
Triple n: <Keio University> </location/location/containedby><Japan>
Answer n:
Output:
1. Keio University is located within the United States of America.
2. Keio University contained by the United States of America.
. . . . . .

For each triples that answers with fewer than 5 sentences, we will call GPT4's API again with the following prompt to further enrich it. In order to improve the quality and richness of the answers, we will include a more detailed prompt in this repeated API call. In this way, GPT4 can further explore and integrate relevant information based on new prompts, generating richer and more detailed answers. This answer not only includes more sentences and details, but also better demonstrates the knowledge domain and contextual environment involved in triplets.

Input:
Task definition: Could you give me some synonymous description for this sentence:
Templates:
Sentence 1:Boston College is located within the United States of America. Answer 1:
1. Boston College is contained by the United States of America.
2. Boston College is situated in the United States of America.
. . . . . .
Sentence n: Ithaca College is located in New York.
Answer n:
Output:
1. Ithaca College is situated in New York.
2. The location of Ithaca College is New York.
. . . . . .

A.1.2 Step 2 - Writing Diverse Questions

In this step, we have obtained world knowledge presented in natural sentence form from the previous step. To ensure the diversity of the final multi round dialogue[42], we generated question and answer pairs (QA pairs) for each natural sentence using prompts in the main text. The final generated question and answer pairs will have rich content and diverse forms, covering various possible dialogue scenarios and topics. These question and answer pairs can be used as materials for constructing multiple rounds of dialogue and as datasets for training dialogue models, improving the generation ability and interaction experience of dialogue models. These QA-pairs are in the following format.

```
{"Q": "What is the form of government in the Dominican Republic?",
"A": "The Dominican Republic has a form of government that is a republic."}
{"Q": "Where is the location of Ithaca College?",
"A": "Ithaca College is located in New York."}
...
```

A.1.3 Step 3 - Constructing Multi-Turn Conversations

To construct engaging and informative multi-turn conversations that effectively promote learning for each entity, we embark on a meticulous process. Initially, we gather all the question-answer (QA) pairs related to each entity, ensuring a comprehensive understanding of the subject matter. The heart of our approach lies in the careful selection and composition of questions. We randomly select a batch of questions from the QA pairs set of a particular entity. This ensures that the questions cover a wide range of topics related to the entity, making the conversation diverse and engaging.To further enrich the conversational experience, we introduce an innovative element: the random replacement of entity concepts with corresponding images. With a 50% probability, we substitute the textual description of an entity with a visually arresting image. This not only adds a visual dimension to the conversation but also aids in the comprehension of abstract or complex concepts. Moreover, this approach ensures that each entity is represented in a unique and engaging manner, fostering adequate learning and understanding. By combining textual and visual elements, we create a immersive and interactive learning experience that is tailored to the needs and preferences of the learners.

A.2 Image Filter for Constructed Data

For each entity in the meticulously constructed dataset IWK300K, we have the capability to retrieve the original images associated with it. On average, each entity boasts approximately 61.23 images, providing a rich visual representation of the entity's diverse attributes and characteristics. However, upon closer inspection, we observe that a significant portion of these images contain excessive irrelevant information and are highly disconnected from the semantic essence of the entity. To ensure that Large Vision-Language Models (LVLMs) can effectively learn from the images of these entities, it becomes imperative to meticulously filter out these "dirty" images. This filtering process not only enhances the quality of the dataset but also ensures that the LVLMs can focus on extracting meaningful and relevant features from the images. As Illustrated in Figure 3, the process of selecting an appropriate image from the Multimedia Knowledge Graph (MMKG) involves three distinct steps: 1) Filtering Gate, 2) Semantic Gate and 3) Refining Gate. The following is the detailed introduction for these three steps.

A.2.1 Filtering Gate

In this gate, we filter images from an image perspective by computing similarity scores between the images of an entity. We utilize two algorithms, dHash and CLIP, to extract features and calculate similarity scores. Specifically, for an image set of one entity, IMG=$(img_1, \ldots, img_n)$, where $img_{(.)}$ represents a candidate image of a entity. Then, for each candidate image, we compute its similarity to the other images and sum these similarities. The dHash algorithm can be computed using the following formula.

$$S_{dHash}^{img_i} = \sum_{t=1, t!=i}^{n} matmul(dHash(img_i), dHash(img_t)^T). \tag{6}$$

Similarly, the CLIP algorithm could be computed with the following formula.

$$S_{CLIP}^{img_i} = \sum_{t=1,t!=i}^{n} matmul(CLIP(img_i), CLIP(img_t)^T). \tag{7}$$

Through the calculations of the above two algorithms, we could find the image set exists the situation where two images are identical. To ensure that each selected image only appears once and avoid the high ranking of weighted similarity caused by duplicate images. We retain only one of every two images with a similarity of 1 in two algorithms, and filter out the rest. Then we can obtain the final score by normalizing the mean of the two similarities with the following formula, where the λ is set to 0.2.

$$S_{final}^{img_i} = \lambda S_{dHash}^{img_i} + (1-\lambda) S_{CLIP}^{img_i}. \tag{8}$$

Finally, our filter gate utilize the $S_{final}^{img_i}$ as the score for sorting and take the top 10 image for the next step filtering.

A.2.2 Semantic Gate

In semantic gate, we filter images from its corresponding entity texts perspective. For this purpose, we use CLIP to obtain text and image features, and we compute the similarity between text and image features to represent the semantic similarity. CLIP is a multi-modal pre-trained model that uses a dual-stream encoding framework and has trained on large-scale data comprising four hundred million image and text pairs collected from the internet. It is highly effective on a wide range of downstream tasks and possesses a powerful ability to project text and image into the same semantic space.

Specifically, given a entity text t and its first step filtered images set $IMG = (img_1, \ldots, img_{10})$, we would encode entity text t by CLIP text encoder while encode image $img_{(.)}$ by CLIP visual encoder. The similarity between entity text t and its candidate image img_i could be computed in the following formula.

$$S_{semantic}^{t-img_i} = matmul(CLIP_{visual}(img_i), CLIP_{text}(t)^T). \tag{9}$$

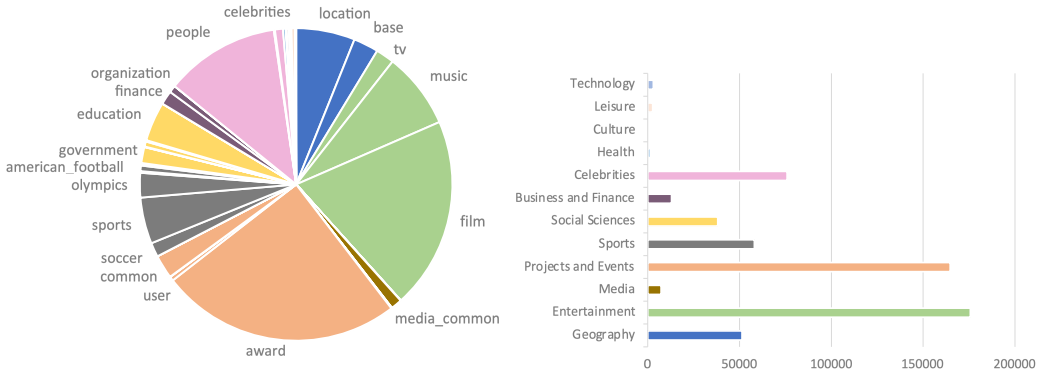
Then we could obtain the order of candidate image set based on the calculated semantic similarity score and take top 5 image for the last step selecting.

A.2.3 Refining Gate

On the account of the limited filter ability of existing model or algorithm, we utilize human experts to select a image from the previous step. Given a entity text t, its filtered image set $IMG = (img_1, \ldots, img_5)$, the corresponding set of constructed natural sentence (Appx 1.1) and corresponding Wikipedia, human experts would firstly select image based on their understanding of entity text and its constructed single turn conversations. (e.g. entity "20th century fox" and its constructed natural sentence "The flim xxxx is released by the studio 20th century fox"). There are 5 human experts in total. Each human expert will select an appropriate image and finally count the number of times each image has been selected. When the number of times is greater than or equal to 2, this image will become the final selection result. For most entities in the dataset, human experts will have a consistent final result. The rest of entities will be combined to Wikipedia, origin images set and Google search engine to find a appropriate image.

B Dataset Diversity

To assess the diversity of sentences within the IWK-300k datasets, we conducted a topic frequency analysis of the world knowledge encompassed in the IWK-300k corpus. As depicted in Figure 9, we initially quantified the overall frequency of each topic across the entities present. Notably, we identified a total of 63 distinct topic types. Subsequently, through a rigorous process of human classification and summarization, these topics were grouped into 12 distinct areas, collectively covering a broad and diverse range of real-world domains.



(a) (b)

Figure 9: The distribution of topics in our dataset.

Example1 JSON of our IWK-300k dataset


```
{
    "id": "m0291hr0",
    "conversations": [
      {
        "from": "human",
        "value": "Who␣is␣the␣director␣of␣this␣film?␣\n<image>"
        "image_list": [m.0291hr.jpg]
      },
      {
        "from": "model",
        "value": "The␣film␣\"High Anxiety\"␣was␣produced␣by␣Mel␣Brooks."
        "image_list": []
      },
      {
        "from": "human",
        "value": "What␣studio␣handled␣the␣distribution␣of␣the␣film␣\"High
            Anxiety\"␣?␣Could␣you␣give␣a␣picture␣of␣it?␣"
        "image_list": []
      },
      {
        "from": "model",
        "value": "The␣film␣\"High Anxiety\"␣was␣distributed␣by␣the␣studio
            ␣20th␣Century␣Fox.␣Here␣is␣a␣picture␣of␣it:␣\n<image>"
        "image_list": ["m.016tt2.jpg"]
      },
      {
        "from": "human",
        "value": "Do␣you␣have␣any␣similar␣moments?"
      },
      {
        "from": "model",
        "value": "20th␣Century␣Fox␣had␣distributed␣the␣film␣\"The
            Happening.\""
        "image_list": []
      }
    ]
}
```

Example2 JSON of our IWK-300k dataset


```
{
    "id": "m01_x6v0",
    "conversations": [
      {
        "from": "human",
        "value": "Which␣university␣did␣<image>␣graduate␣from?"
        "image_list": ["m.01_x6v.jpg"]
      },
      {
        "from": "model",
        "value": "Trey␣Parker␣was␣educated␣at␣Berklee␣College␣of␣Music."
        "image_list": []
      },
      {
        "from": "human",
        "value": "Where␣is␣the␣Berklee␣College␣of␣Music␣situated?␣Could␣
            you␣give␣a␣picture␣of␣it?␣"
        "image_list": []
      },
      {
        "from": "model",
        "value": "The␣Berklee␣College␣of␣Music␣is␣located␣in␣Boston.␣Here
            ␣is␣a␣picture␣of␣it:␣\n<image>"
        "image_list": ["m.02g839.jpg"]
      }
    ]
}
```


C Case study

The four formats of the IWK-300K dataset endow KnowBLIP with a diverse array of functionalities, enabling it to handle a wide range of tasks and scenarios related to real world knowledge. To illustrate these capabilities, we present several concrete user cases in Figure 11-15, each showcasing a unique aspect of KnowBLIP's versatility.

C.1 Text Perception and Generation

One core innovation of KnowBLIP lies in its remarkable ability to perceive and generate real-world knowledge based on the contextual nuances of user conversations. This aspect is particularly valuable in open-world scenarios, where individuals often have intricate and multifaceted ideas about the world, transcending mere general knowledge. Consider a conversation between a user and KnowBLIP. The user might express an interest in learning about a particular topic, say, the film "High Anxiety". Our KnowBLIP with its unique capability, not only understands the user's intention but also responds with accurate real-world knowledge related to the entity "High Anxiety." It might offer details about the film's director, plot and so on. Moreover, KnowBLIP excels in handling interleaved text and images with different entities in multi-turn conversations. Imagine a conversation where the user first asks about "High Anxiety" and then transitions to another entity, such as "20th Century Fox", the studio that produced it. KnowBLIP seamlessly shifts its focus, maintaining coherence and depth throughout the conversation.

C.2 Visual Perception and Generation

The other core creation of KnowBLIP lies in its ability to not only comprehend but also retrieve images related to world knowledge entities based on the contextual conversation flow provided by users. For instance, consider entities like "Richard Rodgers", "New York, I Love You", "University of Oxford", and "Germany". These entities are rich in contextual information and often form the backbone of conversations about various topics. In an open-world scenario, this feature of KnowBLIP

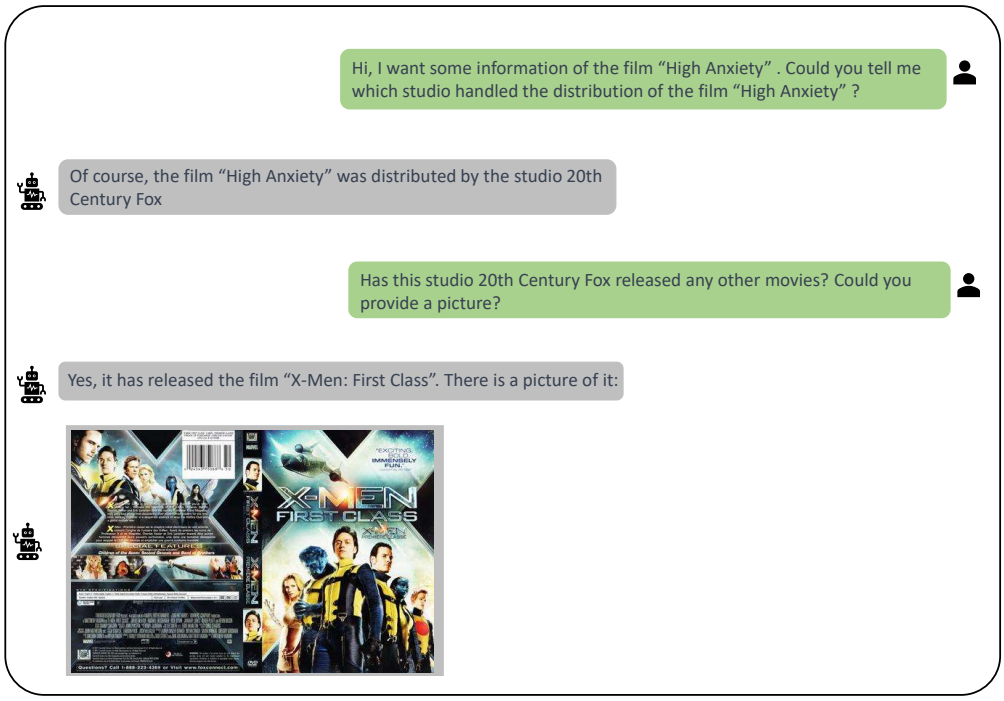


Figure 10: The visualization of KnowBLIP's text understanding, i.e. Text2Text and Text2Image tasks.

becomes particularly handy. Many times, users may simply input an image containing some form of world knowledge and inquire about its associated information. Such inquiries could range from identifying landmarks in a photograph to understanding the historical significance of a painting. Furthermore, there are instances where users actively seek images related to specific world knowledge entities. For example, a user might be interested in exploring images related to the movie “X-Men: First Class” to gain a deeper understanding of its characters, settings, or themes. In such cases, KnowBLIP's ability to retrieve relevant images becomes invaluable. However, it's worth noting that not all image generation schemes are equipped to handle world knowledge effectively. As illustrated in Figure 8 (a), the use of a stable diffusion model for image generation can sometimes fall short in this regard. While the model may produce visually appealing images, it may lack the contextual depth and world knowledge that KnowBLIP is capable of incorporating. To address this gap, KnowBLIP incorporates a robust knowledge base and advanced retrieval mechanisms. This allows it to not only generate images but also ensure that these images are informed by relevant world knowledge. Whether it's retrieving images of famous landmarks or exploring the visual representations of fictional worlds, KnowBLIP provides users with a richer and more informative visual experience.

Figure 11: The visualization of KnowBLIP’s text understanding, i.e. Text2Text and Text2Image tasks.

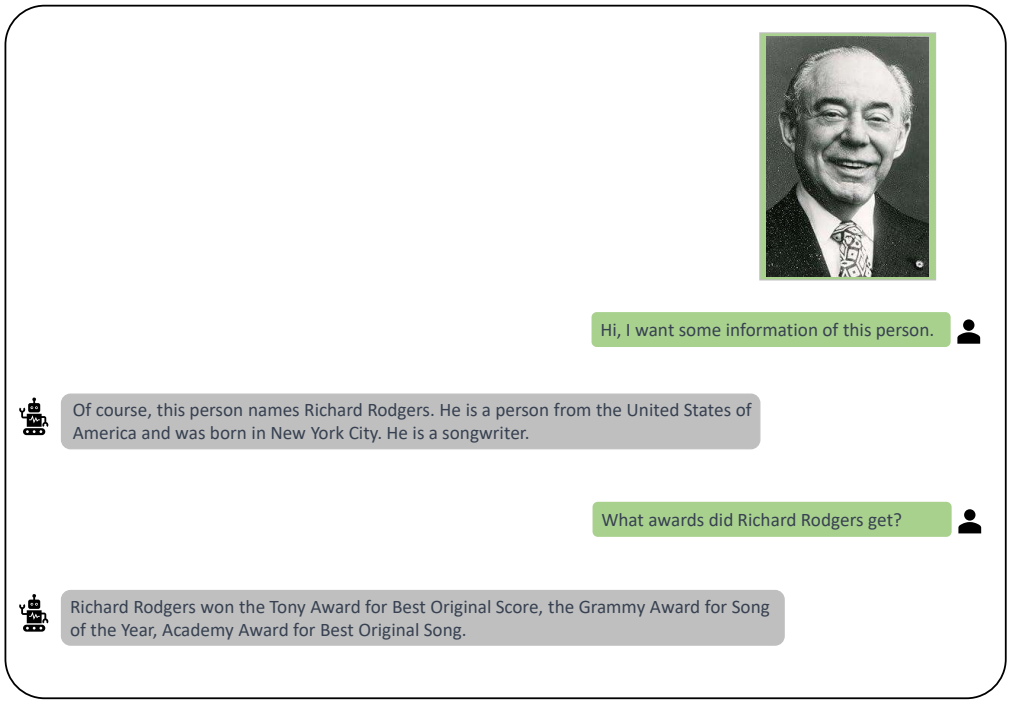


Figure 12: The visualization of KnowBLIP’s image understanding, i.e. Image2Text and Image2Image tasks.

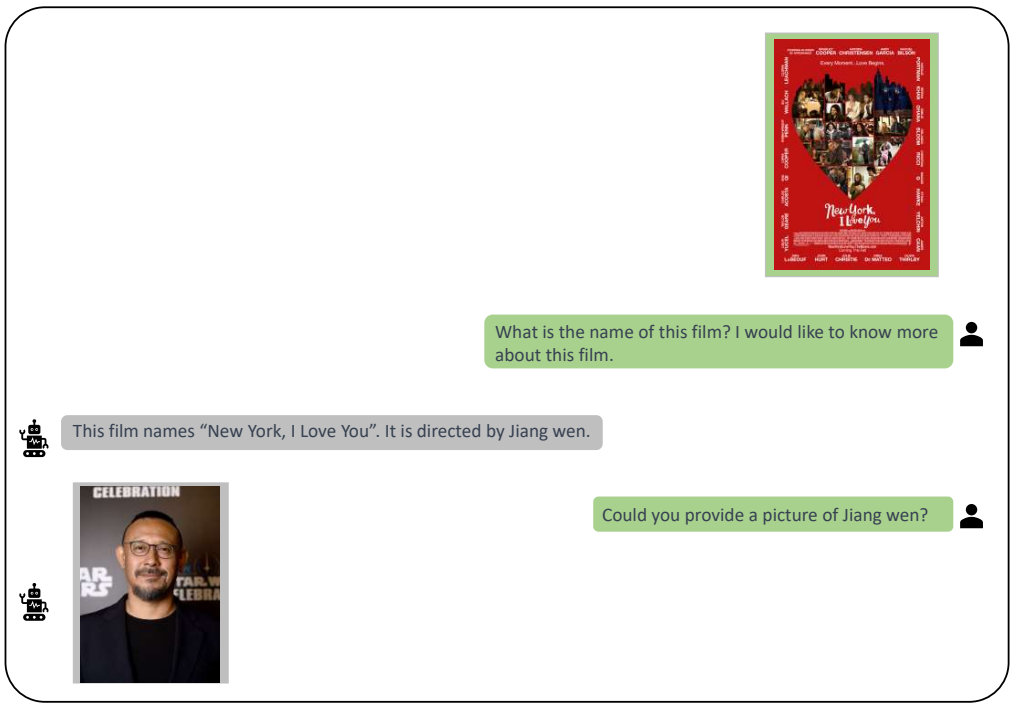


Figure 13: The visualization of KnowBLIP’s image understanding, i.e. Image2Text and Image2Image tasks.

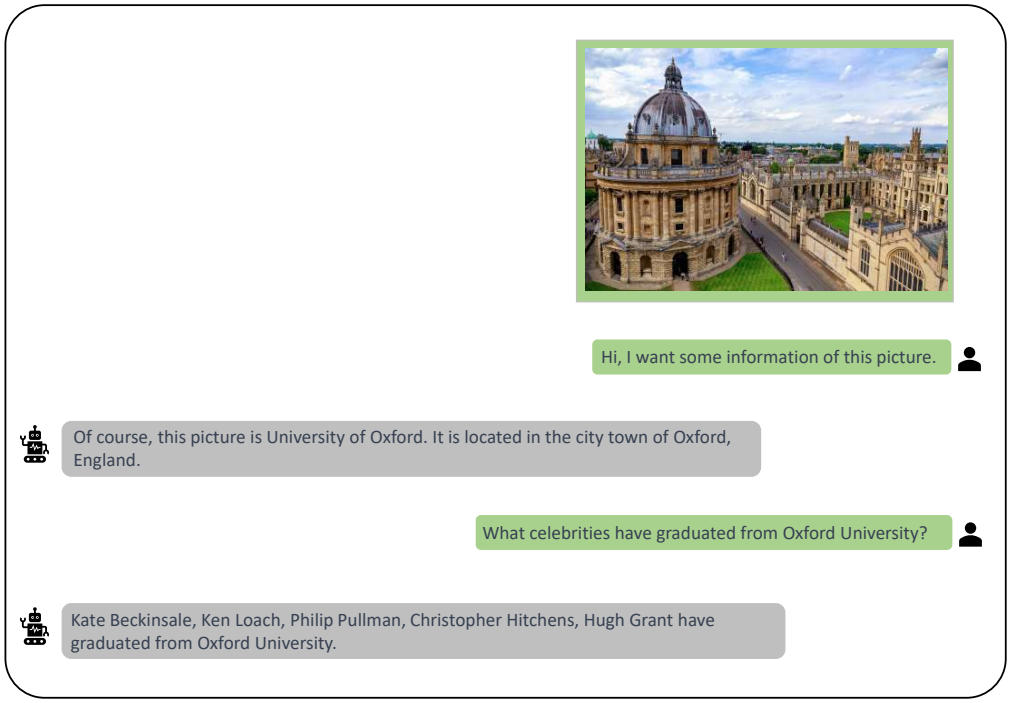


Figure 14: The visualization of KnowBLIP’s image understanding, i.e. Image2Text and Image2Image tasks.

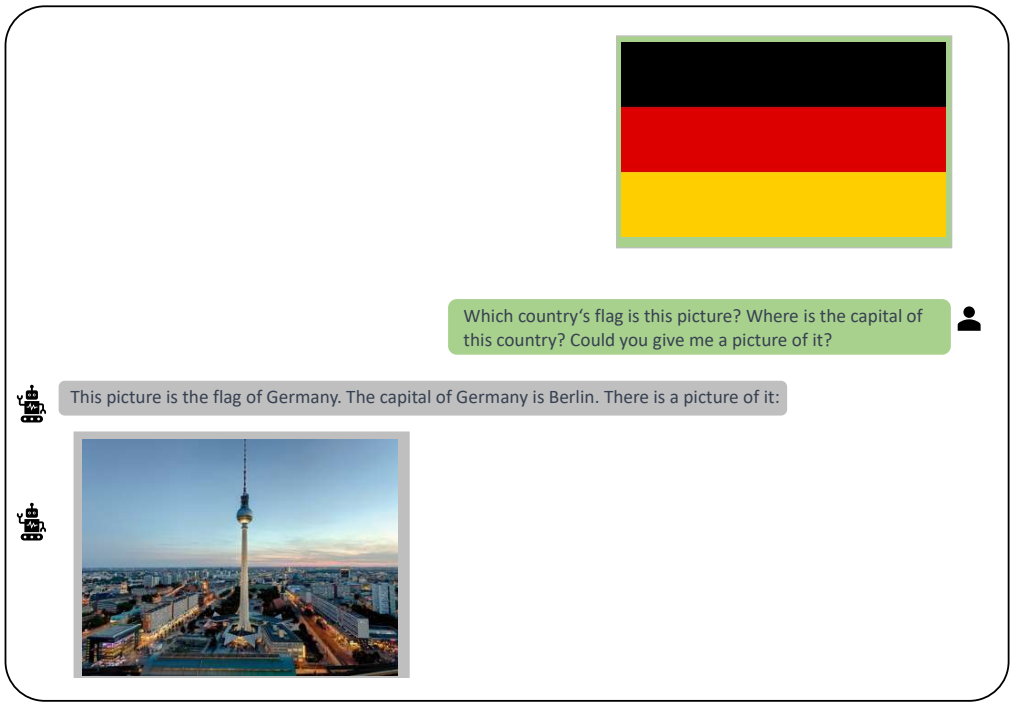


Figure 15: The visualization of KnowBLIP's image understanding, i.e. Image2Text and Image2Image tasks.

D Datasheet of IWK-300k

In this section, we provide a datasheet about our dataset.

D.1 Motivation For Datasheet Creation

Why was the datasheet created? (e.g., was there a specific task in mind? was there a specific gap that needed to be filled?)

This dataset was developed to facilitate research on Multi-modal Large Language models for the world knowledge based on their image-text understanding ability. While there are existing datasets for Multi-modal Large Language models to percieve image and text, none of them include a comprehensive set of image-text pair for learning world knowledge. By creating a multi-turn and interveled annotated dataset, we aim to fill this gap and provide researchers with a resource that can be used to develop more complete ablities multi-modal models for world knowledge.

Has the dataset been used already? If so, where are the results so others can compare (e.g., links to published papers)?

Our dataset is proposed for the first time. This dataset was built upon a subset of existing multi-modal knowledge graph (FB15K[29]), which is consisted by structured triples with the relation in the real world. Our new dataset, IWK-300K, serves as an extension to the this multi-modal knowledge graph dataset, providing more comprehensive world knowledge annotations in conversation format for multi-modal models, which can be used to improve the complete abilities of multi-modal models.

What (other) tasks could the dataset be used for?

This dataset could be used for diverse tasks under world knowledge, including but not limited to tasks such as image caption, VQA, image generation, cross-modal retrieval.

Who funded the creation dataset?

None. We built this dataset at our own expense to promote the development of academia and industry.

Any other comment?

None.

D.2 Datasheet Composition

What are the instances?(that is, examples; e.g., documents, images, people, countries) Are there multiple types of instances? (e.g., movies, users, ratings; people, interactions between them; nodes, edges)

The instances in this dataset are images and triples from the FB15K dataset[43]. The original images and triples contain information about the the image of nodes and relation between nodes. Our new dataset is densely translateed the original images and triples into multi-turn, multi-modal and interleaved conversations established through GPT-4 and human experts, which contains with 4 formats (text-text, image-text, image-image, text-image)

How many instances are there in total (of each type, if appropriate)?

IWK-300k is a dataset of $300,199$ single rounds of conversation, averaging 5.95 rounds per complete multi-turn conversation. Each single turn conversation includes two entity and their relation. The average number of images per entity is 61.23.

What data does each instance consist of ? "Raw" data (e.g., unprocessed text or images)? Features/attributes? Is there a label/target associated with instances? If the instances related to people, are subpopulations identified (e.g., by age, gender, etc.) and what is their distribution?

The IWK-500K dataset consists of multiple types of instance. The types consist of: (a) text to text conversation, (b) text to image conversation, (c) image to text conversation, (d) image to image conversation.

Is there a label or target associated with each instance? If so, please provide a description.

Yes, a single turn conversation often constructed by question of head entity in origin triple, while the label is the response of conversation constructed by tail entity in origin triple.

Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.

The dataset does not include all complete information on a entity, as this information is also not available from the original dataset.

Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.

Not applicable.

Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).

No. IWK-500K only contains entity and its relation between entities on public wikipedia.

Are there recommended data splits (e.g., training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.

Our dataset is split into 3 non-overlapping subsets, where 0.6, 0.2 and 0.2 are used for training, validation and testing. Specifically, for one triplet that has four formats in different formats of question-answer pair, we would split its different formats into training, validation and testing sets, which ensures the world knowledge of this triplet having trained.

Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.

Noise and errors are inevitable in datasets. The most likely source of errors is incorrect labels or relation in the origin dataset or in the semi-automatic data production from GPT-4 caused by its hallucination. However, the IWK-500K dataset collection process has multiple quality assurance steps by human experts that aims to substantially reduce the prevalence of noise and errors.

Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.

Our dataset is developed from the existing FB15K dataset. The FB15k dataset contains knowledge base relation triples and textual mentions of Freebase entity pairs. It has a total of 592,213 triplets with 14,951 entities and 1,345 relationships. The FB15K dataset can be accessed here: https://github.com/mniepert/mmkb/blob/master/FB15K/FB15K_EntityTriples.txt.

Any other comments?

None.

D.3 Collection Process

What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)? How were these mechanisms or procedures validated?

We use GPT-4 API to semi-automatic data production. Prompts had detailed in main content.

How was the data associated with each instance acquired? Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.

A full description appears in the associated paper and its appendix. The three steps include 1) transforming triplets into natural sentences, 2) writing diverse questions, and 3) constructing multi-turn conversations. 4) Human experts review. However, briefly:

- Transforming Triplets into Natural Sentences. Some few shot templates, the prompt and required translated entity is given to GPT-4 to produce data.
- Writing Diverse Questions. generated question and answer pairs (QA pairs) for each natural sentence using prompts given to GPT-4.
- Constructing Multi-Turn Conversations. Randomly select a batch of questions from the QA pairs set of a particular entity
- Human experts review.

If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?

All images from the FB15K had filtered by a sampling strategy, which including three distinct steps: 1) Filtering Gate, 2) Semantic Gate and 3) Refining Gate. In the main content is involved the detailed introduction for these three steps.

Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?

The Artificial Intelligent students from South China Technology of University were involved in reviewed of the data. These students were compensated based on the number of single conversation they reviewed, with payment free.

Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.

According to the original paper [29], the collection of FB15K images took place in 2015. The attribute annotations were conducted in December 2023 - Third 2024.

D.4 Data Preprocessing

Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remainder of the questions in this section.

No preprocessing. The labeling of the dataset was done manually by human annotators.

Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.

Not applicable.

Is the software used to preprocess/clean/label the instances available? If so, please provide a link or other access point.

Not applicable.

Does this dataset collection/processing procedure achieve the motivation for creating the dataset stated in the first section of this datasheet? If not, what are the limitations?

Not applicable.

Any other comments

None.

D.5 Dataset Distribution

How will the dataset be distributed? (e.g., tarball on website, API, GitHub; does the data have a DOI and is it archived redundantly?)

The FB15K dataset can be accessed through the following link: [3]. Additionally, the newly annotated dataset, IWK-500K, is uploaded here: [4].

When will the dataset be released/first distributed? What license (if any) is it distributed under?

The dataset is already available.

Are there any copyrights on the data?

The FB15K in released under BSD-3-Clause License.

Are there any fees or access/export restrictions?

There are no fees. Our dataset is also intended to be used for research and educational purposes.

Any other comments? None.

D.6 Dataset Maintenance

Who is supporting/hosting/maintaining the dataset?

The dataset and the website where the annotations are released will be maintained by the authors of the manuscript.

Will the dataset be updated? If so, how often and by whom?

We do not have concrete plans as of yet; we will announce any updates on the dataset website.

[3] https://github.com/mniepert/mmkb/tree/master

[4] https://drive.google.com/drive/folders/1ZgfApnIhzXg8n3R_kwikix5C0Klv6UP8

How will updates be communicated? (e.g., mailing list, GitHub)

Updates will be communicated through the dataset website: [5].

If the dataset becomes obsolete how will this be communicated?

Through our github website: [6].

Is there a repository to link to any/all papers/systems that use this dataset?

There is a repository, maintained by the authors of the manuscript at [7].

If others want to extend/augment/build on this dataset, is there a mechanism for them to do so? If so, is there a process for tracking/assessing the quality of those contributions. What is the process for communicating/distributing these contributions to users?

The dataset is under the MIT licence so anyone has freedom to do so. Currently, we do not have mechanisms in place; however, others may contact us to discuss potential use cases if they prefer.

[5] https://drive.google.com/drive/folders/1ZgfApnIhzXg8n3R_kwikix5C0Klv6UP8

[6] https://github.com/sdjhshbswp/IWK500K

[7]

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We have summarized the main contributions at the end of the introduction.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have summarized the limitations in the experiment and appendix.

3. **Theory Assumptions and Proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: Our paper does not include theoretical results.

4. **Experimental Result Reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provided the available URL of all codes and datasets.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We have provided the available URL of all codes and datasets.

6. **Experimental Setting/Details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: We have displayed it in the experiment.

7. **Experiment Statistical Significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We will add this experiment in the future.

8. **Experiments Compute Resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We have explained it in the setting.

9. **Code Of Ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

Answer: [Yes]

Justification: We have confirmed the Ethics Guidelines.

10. **Broader Impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We believe our work will inspire the multi-modal field.

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: Our paper poses no such risks.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We have provided the available URL of datasets.

13. **New Assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: Our paper does not release new assets.

14. **Crowdsourcing and Research with Human Subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.

15. **Institutional Review Board (IRB) Approvals or Equivalent for Research with Human Subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: Our paper does not involve crowdsourcing nor research with human subjects.