



# Undergraduate Project Report

## Evaluating the encoding competence of visual language models using uncommon actions

**Author:** Chen Ling

**Affiliation:** Beijing University of Post and Telecommunications

**Email:** 20212127000@bupt.edu.cn

**Supervisor:** Nai Ding\*

**Affiliation:** Zhejiang University

**Email:** ding\_nai@zju.edu.cn

## Table of Contents

<i>Abstract</i> . . . . .	2
<i>Keywords</i> . . . . .	2
<i>Chapter 1: Introduction</i> . . . . .	3
<i>Chapter 2: Background</i> . . . . .	5
2.1 Relation Work . . . . .	5
2.2 Image generation tools . . . . .	5
2.3 Evaluation models . . . . .	6
2.3.1 Vison and Langugae models . . . . .	6
2.3.2 Contrasting Learning Models . . . . .	6
2.4 Fine-tuning method . . . . .	7
2.5 Innovations of this study . . . . .	8
<i>Chapter 3: Design and Implementation</i> . . . . .	9
3.1 Verb library filtering . . . . .	9
3.1.1 Introduction of VerbNet . . . . .	9
3.1.2 Pipeline of constructing Verb library . . . . .	9
3.1.3 Classification of Verb library . . . . .	11
3.2 Uncommon-Text Generation . . . . .	11
3.3 Uncommon Image-Text Construction . . . . .	14
3.3.1 Pipeline of Image generation . . . . .	14
3.3.2 Configuration of image generation . . . . .	16
3.4 VQA dataset Construction . . . . .	17
<i>Chapter 4: Results and Discussion</i> . . . . .	19
4.1 Model selection . . . . .	20
4.2 Result analysis . . . . .	20
4.3 Analysis of comparative learning models . . . . .	22
4.4 Discussion on performance between human and models . . . . .	23
4.5 Evaluating and Training configuration . . . . .	24
4.6 Role of UAIT in promoting semantic understanding . . . . .	24
4.6.1 Stripping away contextual statistical bias . . . . .	25
4.6.2 Higher explainability and diagnosability . . . . .	25
4.7 Implications for further research on multimodal understanding . . . . .	26
4.7.1 Improving the capabilities on modeling semantic role . . . . .	26
4.7.2 Introducing common sense anti-interference mechanism . . . . .	26
4.7.3 Data-driven along with reasoning-driven . . . . .	26
<i>Chapter 5: Conclusion and Further Work</i> . . . . .	28
5.1 Conclusion . . . . .	28
5.2 Reflection . . . . .	28
5.3 Further work . . . . .	29
<i>References</i> . . . . .	31

## Abstract

We propose **UAIT** (**U**ncommon-sense **A**ction **I**mage-**T**ext) dataset, a new evaluation benchmark designed to test the semantic understanding ability of visual language models (VLMs) in uncommon-sense action scenes. Unlike previous datasets that focus on common visual scenes with statistical frequency advantages, UAIT challenges models with grammatically reasonable but semantically counter-common sense image-text pairs. Such tasks require models to go beyond superficial pattern recognition and demonstrate a deep understanding of agent-patient relationships and physical feasibility.

To build UAIT, we designed a semi-automated process to synthesize high-quality uncommon-sense image-text samples using large language models, few-shot prompt engineering, and text-to-image generation. Each sample is accompanied by a carefully designed multiple-choice question to test the model’s competence in fine-grained reasoning. We evaluate multiple state-of-the-art visual language models and compare them with models based on contrastive learning. Experiments show that all models perform significantly worse than humans in semantic judgment, especially in distinguishing grammatical correctness from semantic rationality.

Further experiments show that even the lightweight model can improve its accuracy after fine-tuning, demonstrating the great potential of directional adaptation. However, there is still a significant gap between the model and humans, reflecting the fundamental limitations of current multimodal understanding. This study not only reveals the key weaknesses of VLMs, but also provides diagnostic tools and research directions for the development of robust models with real visual semantic reasoning capabilities.

## Keywords

Vision-Language Models (VLMs), Uncommon-sense Reasoning, Action Recognition

## Chapter 1: Introduction

In recent years, vision and language models (VLMs) have made significant breakthroughs in the field of artificial intelligence, especially in multimodal understanding tasks such as visual question answering (VQA), image description, and visual commonsense reasoning. These models aim to bridge the semantic gap between visual and textual information by jointly encoding features of the two. Although state-of-the-art models have shown impressive performance in a wide range of applications, their true understanding of fine-grained visual semantics, especially those involving action reasoning and counter-common sense reasoning, remains under-explored.

When people see an uncommon action image, they can quickly identify the unusual, counter-logical, and incongruous parts of it. Take the example in Figure 1: a rabbit is dragging a tiger. Although the components in the picture are all very normal objects in real life, the combination relationship between them appears unusual. Humans can easily recognize the unusual in pictures, but this is preceded by a multi-step reasoning process and requires common sense problems that rely on daily experience, physical knowledge, or various social and cultural norms.

In this work, we present **UAIT**(Uncommon-sense Action Image-Text Dataset), which aims to systematically evaluate the encoding competence of VLMs when facing visual scenes that exhibit unusual or counter-intuitive actions. Specifically, we want to determine whether the current state-of-the-art models can correctly interpret the semantics of images in situations where the actions violate common sense or challenge existing assumptions (such as "zebra attacks lion"). Unlike traditional tasks that focus on evaluating the recognition of common patterns and concepts, our focus shifts to edge cases that require true visual and textual understanding rather than relying solely on statistical biases or memory associations.

To this end, we propose and implement a novel pipeline for building the UAIT - a curated benchmark consisting of synthetically generated visual scenes and their corresponding textual descriptions. The scenes are selected from semantically sensitive verbs in VerbNet such as those whose meaning depends heavily on the agent-patient relationship, and are carefully designed to simulate plausible yet unusual events. The project synthesizes richly annotated data using large language models and text-to-image generation techniques such as Stable Diffusion. We then use this dataset to evaluate the performance of various VLMs in distinguishing semantically opposing interpretations.

The technical core of the work includes:

- *Verb selection principles* based on syntactic reversibility and visual clarity.
- Methods for generating syntactically correct but semantically misleading sentences



### Why is the action scene uncommon ?

- The tiger and the rabbit are the relationship of prey and hunting**
- The power of the rabbit is far smaller than that of the tiger**

Figure 1: This is an example from UAIT, the correct description of the action scene is “a rabbit is dragging a tiger” . Below are the reasons why the image is beyond the common-sense understanding.

using *prompt engineering* and *few-shot* learning techniques.

- *Text-to-image generation* for unusual visual events guided by detailed scene descriptions.
- *Evaluation* through accuracy comparison and reasoning analysis.

Our work also provides an evaluation framework to measure the deep understanding of visual actions by current VLMs, rather than just superficial pattern matching. The experimental data on UAIT shows that the performance of state-of-the-art visual language models lags behind humans on VQA tasks. Our findings show that SOTA models have difficulty distinguishing syntactic correctness from semantic accuracy in action scenes, which highlights a key flaw in current multimodal learning systems.

In summary, we built a novel action-oriented image-text dataset, designed and verified a complete generation pipeline, and started benchmarking multiple VLMs. Our results show that even state-of-the-art visual-language models perform poorly on the challenging UAIT benchmark. This result aims to advance our understanding of the limitations of visual-language models and motivate more robust approaches to semantic encoding of visual actions.

## Chapter 2: Background

With the development of large-scale pre-training technology, vision-and-language models (VLMs) have made significant progress. Models such as GPT-4(1), LLaVA(2), and Qwen2-VL(3) are representatives of this trend, and they have demonstrated strong capabilities in multimodal tasks. However, despite the outstanding performance of these models, current vision-language models often rely heavily on statistical regularities and contextual priors. For example, in real-world datasets, some visual-text pairs are more common than their abnormal counterparts. This frequency bias can cause the model to make overconfident but wrong judgments when encountering visual inputs that violate common sense.

### 2.1 Relation Work

Currently popular datasets such as COCO-A(4) (Common Objects in Context - Actions) attempt to enrich image data through action annotations, but such datasets are still dominated by frequently occurring action-object combinations, and only provide limited examples of inverted or ambiguous agents and recipients, making it difficult to effectively evaluate the true reasoning ability of the model.

Winoground(5) is a benchmark designed to evaluate the cross-modal composition ability and fine-grained semantic understanding of visual language models. It constructs semantically very similar but slightly different image-text pairs to examine whether the model can accurately match images with their descriptions. However, Winoground mainly focuses on common scenes in life and does not contain counterintuitive examples, which limits its ability to test the deep understanding of the model.

### 2.2 Image generation tools

To address the lack of diversity and unusual action scenes in current datasets, we introduced Stable Diffusion, an advanced text-to-image generation model based on a diffusion process.

Stable Diffusion(6) is a generative artificial intelligence technology based on a diffusion model. It is mainly used to convert text descriptions into high-quality images. Its core idea is to gradually convert the initial random noise into an image that conforms to the semantics of the input text through a step-by-step "denoising" process. In the training phase, the model uses a large amount of image and text data to build a mapping relationship between text and visual features, and compresses the image data into the latent space for processing, which greatly reduces the consumption of computing resources and

improves the generation efficiency and image resolution. Its workflow mainly relies on a back-diffusion process, which starts with a noisy image and gradually extracts and reconstructs detailed image content in multiple iterations. Also, the text encoder based on the Transformer(7) architecture is used to accurately capture the semantic information in the input description, so that the final generated image can accurately reflect the creativity and details contained in the text. Due to its open source and flexible and customizable characteristics, Stable Diffusion not only promotes the integrated application of computer vision and natural language processing, but also provides strong technical support for a variety of scenarios such as artistic creation, advertising design, virtual reality.

In our work, Stable Diffusion is used to synthesize vivid, clear, and visually coherent unusual action images, which is an important tool for building the images of UAIT.

## 2.3 Evaluation models

### 2.3.1 Vision and Language models

We used multimodal models such as LLaVA(2) (Large Language and Vision Assistant) for evaluation.

LLaVA (Large Language and Vision Assistant) is a multimodal interaction model that combines vision and language processing capabilities, aiming to achieve efficient interconnection between image and text information. The model uses a pre-trained large language model as a basis, and combines it with a visual encoder to extract image features, so that the model can understand and generate descriptions and answers related to the image content. During the training process, LLaVA achieves deep alignment between visual features and language expressions by fine-tuning on a large number of image and text corresponding datasets, so that it can accurately capture image details and provide targeted text feedback in conversations. Its advantage is that it can support mixed image and text input, which can not only complete traditional text question-answering tasks, but also interpret, reason and creatively describe image content to meet a variety of practical application needs, such as intelligent customer service, auxiliary diagnosis, content creation and other fields.

In our study, LLaVA and similar models were used to evaluate and test the UAIT dataset.

### 2.3.2 Contrasting Learning Models

CLIP(8) (Contrastive Language-Image Pretraining) is a multimodal model proposed by OpenAI. It uses contrastive learning methods to jointly train on massive image and text pair data, and realizes the alignment of visual and language information in the same

embedding space. The model usually consists of two independent encoders: an image encoder (usually based on a convolutional neural network or a visual Transformer(9) architecture) and a text encoder (based on a Transformer architecture), which extract features of images and text respectively. During the training process, CLIP uses a contrastive loss function to make matching image and text pairs closer in high-dimensional space, while mismatched pairs are farther away. This mechanism enables the model to learn to capture the deep semantic association between images and text. Thanks to the support of massive Internet data, CLIP has demonstrated excellent zero-shot learning capabilities, that is, when faced with unseen categories, it can still accurately identify and classify through natural language descriptions. In addition to its outstanding performance in traditional computer vision tasks such as image classification and retrieval, CLIP’s cross-modal characteristics also provide new ideas and technical support for fields such as image generation, automatic annotation, and human-computer interaction.

We used CLIP and RNN(10)-based RWKV-CLIP(11) as evaluation models for image-text matching related tasks.

## 2.4 Fine-tuning method

Fine-tuning is a commonly used technique in the field of machine learning and deep learning. Its main purpose is to make the model better adapt to the target task by retraining it on a small-scale dataset for a specific task or a specific field based on an existing pre-trained model. Pre-trained models are usually trained on massive data and learn a large number of common features and patterns. Fine-tuning uses this existing knowledge and then conducts additional training in a shorter period of time to make the model parameters better capture the subtle differences in the target field while retaining the original generalization ability. During the fine-tuning process, researchers can choose to freeze some low-level parameters of the model and only adjust high-level parameters. They can also train the entire model, but usually a lower learning rate is used to prevent the destruction of the original knowledge structure. This approach can not only significantly shorten the training time and reduce the dependence on the size of the data, but also help alleviate the overfitting problem because the pre-trained model already has rich representation capabilities.

In our work, we used LoRA(12) as a fine-tuning technology due to the limitation of the computational memory.

## 2.5 Innovations of this study

**Unusual action focus:** We constructed action scenes that deliberately violate common sense expectations, exposing the model to challenges beyond its statistical "comfort zone".  
**Verb-level semantic control:** We screened verbs based on their semantic reversibility (e.g, exchanging the agent and the patient will cause the sentence meaning to change), ensuring that each example tests the model's reasoning ability rather than superficial image recognition.

**Synthetic dataset construction:** Unlike previous datasets that rely on web crawlers or real image annotations, we use large language models and text-image diffusion models to generate data, avoiding data scarcity and copyright issues.

**Fine-grained evaluation mechanism:** Each image in the UAIT dataset is accompanied by a carefully designed multiple-choice question to test the model's ability to understand action role dynamics (agent-patient relationships).

While existing research has laid the foundation for understanding the behavior of multimodal models, our project further advances the field by focusing on an overlooked but critical ability – accurate action reasoning under conditions of semantic role reversal. This allows us to have a deeper and more diagnostic understanding of what visual language models can "see" and how much they really "understand" when interpreting the visual world.

## Chapter 3: Design and Implementation

### 3.1 Verb library filtering

#### 3.1.1 Introduction of VerbNet

Our work started to build UAIT by collecting verbal literals. The action verbs were extracted from verb library – **VerbNet**.

VerbNet is a detailed grammatical and semantic database for English verbs. It is based on the theoretical research results of *Beth Levin(13)* on the grammatical behavior of English verbs. It systematically classifies and finely annotates verbs, revealing the intrinsic connection between different verbs in syntactic structure and semantic role. This resource divides a large number of verbs into multiple levels of categories according to shared semantic features and syntactic characteristics. The verbs in each category are not only similar in basic meaning, but also show consistent grammatical behavior and semantic patterns in actual use, thus providing a solid theoretical foundation for computers to automatically understand and generate natural language. VerbNet records in detail the possible syntactic structure, common collocation components and semantic roles played by each component, such as agent, patient, instrument and result, of each verb, helping researchers and application systems to identify the logical relationship and mechanism of action between the various components in a sentence. In addition, VerbNet is closely related to other lexical resources. It not only makes up for the shortcomings of a single dictionary in describing verb polysemy and grammatical transformation, but also provides rich data support and flexible application interfaces for tasks in the field of natural language processing, such as semantic parsing, machine translation, and information extraction. Through its hierarchical and structured design, VerbNet has become an important tool for exploring the internal structure of verbs, understanding the expression of actions and the organization of events in language, and has played a bridging role in the in-depth study of linguistic theories and the construction of computational models. Its detailed grammatical and semantic annotations provide researchers with an efficient and systematic research platform, which has greatly promoted the development of natural language processing technology and cognitive linguistics.

#### 3.1.2 Pipeline of constructing Verb library

Leveraging the structured semantic resources and construction format provided by VerbNet, we screened and extracted relevant verbs to ultimately establish the **sensitive-action** verb library for visual scene understanding.

In the process of building **sensitive-action**, we set four clear principles:

- **Semantic consistency:** verbs should describe the action relationship between objects in detail, and the classification criteria and definition of this action are clear (such as the action of "salute" must be related to hand and eye movements). In the process of verb screening, we have at least two annotators check the selected verbs to ensure that different people have a consistent understanding of the same action.
- **Semantic non-interchangeability:** We screen out verbs that can reflect the interaction between objects. Specifically, such verbs should be able to clearly describe the semantic structure of "agent-action-patient"(NP: Agent; VERB; NP: Recipient). When the subject and object of such verbs are swapped, the overall meaning of the sentence will also change. For example, the semantics of "a cat chases a mouse" and "a mouse chases a cat" are obviously different, indicating that the verb "chase" clearly expresses directional actions, so it is included in the verb library. However, "a girl is playing with a dog" and "a dog is playing with a girl" have basically the same meaning, which means that "play with" has no clear directional semantics, so it is not suitable for our semantic analysis needs aimed at visual action recognition.
- **Static recognizability:** The selected verbs should be recognizable and understandable through static images rather than dynamic videos. In other words, the selected verbs must meet the definition of visual action, that is, their action characteristics have clear and perceptible performance in one frame of image, and do not need to rely on time dimension information (such as inter-frame difference) or semantic and image information between contexts to be understood by people. For example, actions such as "run" and "kick" can usually be captured in images, while internal cognitive verbs such as "think" are difficult to accurately identify through a single image, so they are not within the scope of selection.
- **Commonness and non-professionalism:** In order to ensure that the verb library has wide applicability, we give priority to common verbs with strong versatility and clear semantics when screening, and avoid selecting terms that are limited to specific fields. For example, "on base" is a professional term in baseball, and "block" also has specific semantics in basketball. Although these words have action meanings in specific contexts, they are difficult to be generally applied in general image semantic analysis. Therefore, such verbs with strong domain characteristics are excluded from the candidate range.

Table 1: classification and examples of the verb library

Class	Verb	Judgement Basis	example
6	40	direction of agent	chase
19	106	active, passive, body contact	crash
12	80	active, passive, tool using	cut
12	80	active, passive, body part	blow
4	21	position of agent	sit

### 3.1.3 Classification of Verb library

Following the principles, we built **sensitive-action** verb library from VerbNet, which contains **53** classes and **318** verbs in total. Table 1 shows the classification of verbs and the basis for image judgment.

The scenes are divided as follows:

The first type of scene: the subject and object of the sentence are distinguished by the direction of the object, such as "chase" and "follow". In this type of scene, the objects in the picture must maintain a consistent direction, and the object is usually placed before the subject.

The second type of scene: the subject and object are determined based on active and passive body contact, such as "punch" and "hit", and the subject in the picture must actively contact the object.

The third type of scene: the subject and object are distinguished by whether tools are used and active and passive action methods, such as "cut someone's hair", in which the subject should contact the object by actively using tools.

The fourth type of scene: the subject and object are divided according to whether there is an action on the body part, such as "bow" and "salute". Although there may be no object contact or tool use in the picture, the movement of a certain object's body part is sufficient to indicate that it is the initiator of the action.

## 3.2 Uncommon-Text Generation

At the stage of constructing Uncommon-Text in UAIT, we used the semantic non-interchangeability of all verbs in the verb library as a theoretical guide. Specifically, we believe that a sentence using a standard subject-verb-object structure (agent+verb+patient) usually presents common-sense semantics, and such sentences are extremely common in everyday language. However, due to the semantic non-interchangeability of verbs themselves, once the two subjects in the sentence - agent and patient - are swapped (patient+verb+agent), the overall semantics of the sentence will change significantly, presenting an anti-common

Table 2: A demonstration example of generating uncommon-sense texts. In *common-sense texts generation*, we use *qwen2(14)* as the large language model assistant. The *Instruct* is a well-designed prompt, which aims to generate a large amount of common sense texts that appear in the language model training corpus, and the content of *assistant* is the answer of the language model. In *uncommon examples construction*, some texts in the answer are selected and the texts are constructed according to the rules in transfer rules.

---

***Common-sense texts generation***

---

**Instruct:**

give me 100 common sentences around the verb 'follow', showing the visual interaction between characters, which can be generated into images by stable diffusion models, the subject and object of the sentence should be human or animals or both, the sentences should be common and familiar.

**Assistant:**

Here are 100 visually descriptive and familiar sentences using the verb "follow", where the subject and object are humans, animals, or both, and the interaction is clearly visual —perfect for generating images using tools like Stable Diffusion:

A student follows a senior across the campus.  
A girl follows her big brother on a bike.  
A blind man follows a guide dog.  
A calf follows a cow into the meadow.  
A chick follows a hen through the yard.  
A man follows his date through a museum.  
A reporter follows a celebrity out of a building.  
A policeman follows a suspect.  
A man follows a vision through a snowy forest.  
A boy follows a butterfly in a field of flowers.  
.....

---

***uncommon examples construction***

---

**original texts:**

a <blind man> follows a <guide dog>  
a <policeman> follows a <suspect>

**transfer rules:**

"<Agent> + verb + <Patient>" → "<Patient> + verb + <Agent>"

**constructed texts:**

a <guide dog> follows a <blind man>  
a <suspect> follows a <policeman>

---

sense or atypical meaning. Therefore, we define this sentence produced after subject swap as an uncommon-sense sentence as the core concept of constructing Uncommon-Text.

On this basis, based on the research results of few-shot learning(15), we realized that

the application of few-shot learning in large language models plays a vital role. This method enables the model to quickly understand and adapt to new tasks through a very small number of examples, without the need for large-scale, specialized training for each task. Few shot learning not only significantly improves the flexibility and efficiency of the model in diverse tasks, but also ensures that the model achieves excellent performance when labeled data is scarce. This provides technical support and theoretical basis for us to use large language models to generate counterintuitive sentences.

In order to efficiently collect a large number of counter-common sense sentences, we designed and implemented a semi-automatic method that combines few-shot learning with large language models to assist in text generation. The specific process is shown in Table 2. Firstly, based on the grouping information in the sensitive-action library, we let the large language model generate semantically common and common-sense sentences for the verbs in each group. This process ensures that the generated basic sentences have clear common sense and structural stability. Then, we preliminarily screened the generated sentences and manually intervened to swap the subject and object in the sentence to construct a small amount of preliminary uncommon-text. In this way, the originally ordinary sentences are given new semantics, reflecting obvious counter-common sense characteristics.

Next, we use these initially constructed uncommon-texts as few-shot examples, input them into the large language model again, and explicitly prompt it to generate a large number of counterintuitive sentences related to such verbs. With the advantages of few-shot learning, the model can capture the specific semantic features and structural patterns required for counterintuitive sentences in a short period of time, thereby achieving large-scale generation. However, although the model learned through few-shot learning already has a certain generation capability, in practical applications, it does not always produce sentences that fully meet the requirements. Therefore, we continue to introduce manual screening and correction links in the generation process, carefully eliminating sentences that fail to fully reflect counterintuitive features or have other semantic deviations.

Finally, we were able to build a high-quality Uncommon-Text dataset through this rigorous process of combining automatic generation with manual review. Each sentence in this dataset has been repeatedly verified and optimized to ensure that it can significantly deviate from common sense in semantics while having a clear and explicit structure and descriptive details. Such a dataset not only enriches the connotation of UAIT, but also provides a solid data foundation for the subsequent evaluation of visual language models, demonstrating an effective method for generating counterintuitive content using a small number of examples and a large language model.

Table 3: example of generating uncommon-text and detailed visual descriptions of the text. In the process of *visual description generation*, *major elements* refer to the elements required to describe the visual description.

---

***uncommon texts generation***

---

**Instruct:**

give me uncommon-sense sentences around the verb 'follow', showing the visual interaction between characters, which can be generated into images by stable diffusion models, the subject and object of the sentence should be human or animals or both, the sentences should be non-common-sense to erase the common-sense bias, for example:

a guide dog follows a blind man.

a suspect follows a police man.

**Assistant:**

Here are the uncommon-sense sentences using the verb "follow", where humans or animals visually interact through the act of following. Each sentence is crafted to be visually distinct and unexpected, ideal for generating imaginative images using Stable Diffusion or similar tools:

A lion in a tuxedo follows a ballet dancer across the stage.

A toddler follows a mime who's floating mid-air with balloons.

A penguin follows a scuba diver through a library.

An astronaut follows a cat walking on the ceiling.

A knight follows a unicorn in a subway tunnel.

.....

---

***visual description generation***

---

**original texts:** a <policeman> follows a <suspect>

**major elements:** *background description, necessary tools or body contact, active or passive*

**constructed texts:** In a busy urban setting, a uniformed police officer walks down the sidewalk unaware, while a man in casual clothes with a wary expression follows closely behind, eyes fixed on the officer.

---

### 3.3 Uncommon Image-Text Construction

#### 3.3.1 Pipeline of Image generation

UAIT was designed to test the cognitive boundaries of current visual language models by using images that require complex common sense reasoning and go beyond the traditional understanding of simple object co-occurrence. The concept of "uncommon" is inherently subjective and its definition varies from person to person. For this reason, we designed a set of methods to collect images from multiple angles and levels to ensure that the final

dataset can reflect the "unconventional" characteristics in a strict sense.

Unlike previous work that mostly focused on reasoning and analyzing existing images, the core concept of UAIT is to use generative technology to create images that are rare or even never seen in real life, a visualized construction example can be seen in Figure 2. Such images can not only break through the limitations of traditional data, but also effectively challenge the ability of visual language models in common sense reasoning and compositional reasoning, especially those scenes that break the conventional expectations of daily experience. Because these images are generated in an innovative way, it can be guaranteed not exist in the pre-training data of most current visual language models to a certain extent, thus providing an independent and rigorous testing platform for model evaluation.

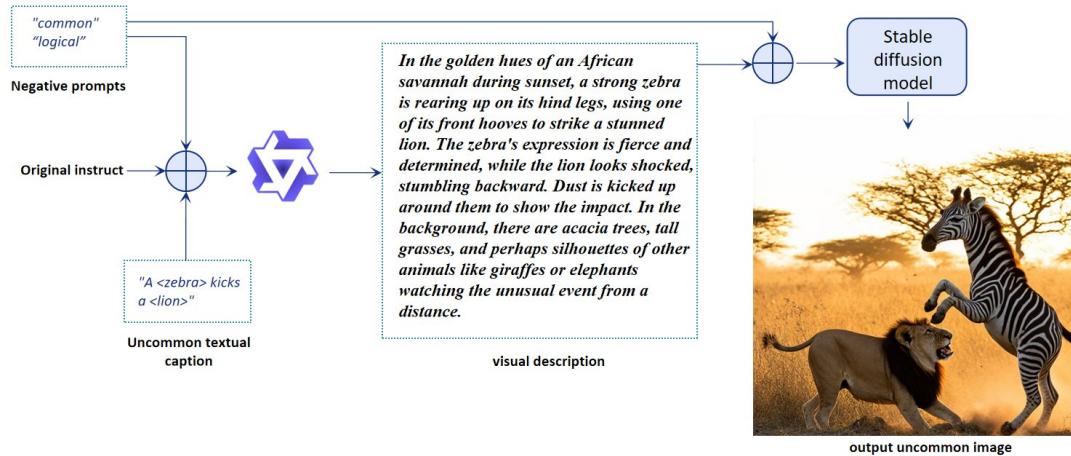


Figure 2: The construction framework of uncommon image-text

To achieve this, we used the advanced text-generated image model, stable diffusion 3. Considering that directly generating "uncommon" images may face the problems of semantic ambiguity and inaccurate description, we introduced an auxiliary generation method based on detailed counter-common sense visual descriptions. In the specific operation process, we first selected and preprocessed a series of representative uncommon-text texts, and input each text into large language model. Through the well-designed prompt engineering, the model can generate a visual language description that has both uncommon sense characteristics and is sufficiently detailed, providing a solid semantic foundation for

subsequent image generation.

After obtaining the preliminary visual description, we did not stop at automatic generation, but supplemented and improved the description content through manual review. The main purpose of this step is to ensure that the description covers detailed objects, scenes, actions, agents and patients, and other key information, so that the subsequently generated images can accurately convey the meaning of "uncommon" both semantically and visually. Next, we input these manually revised detailed descriptions into stable diffusion and generate them in batches of 20 images. During the generation process, we carefully compare and screen each batch of images to ensure that the images can clearly convey the concept of "uncommon" and eliminate those that may cause misunderstandings due to redundant elements. For example, when describing an action scene, we pay special attention to ensure that the identity of the characters in the picture, the logic of the action, and the background scene are all within the scope of common sense, and it is only the specific action part that makes the overall image unusual.

Finally, after the image generation and preliminary screening, we once again comprehensively reviewed and refined the images and their corresponding texts. This stage not only focuses on the accuracy of the facts and the specificity of the description, but also focuses on incorporating the implicit details in the image, such as the specific naming of the subject-verb-object relationship, the clear identification of the tools used, the necessary scenes and background descriptions, etc. After this series of rigorous review processes, we selected the images that best met the standards and finally included them in the uncommon image dataset. The entire process aims to create image data with clear "uncommon" attributes, so that different reviewers can reach a high degree of consensus on the understanding of the image, and it also fully demonstrates our exploration and efforts in the diversity and complexity of images in the process of dataset construction.

### 3.3.2 Configuration of image generation

ComfyUI interface is used to integrate the model and generate images. The detailed model and parameters are as follows:

- **Hardware:** 2×RTX3060 24GB GPUs (NVLink interconnected)
- **Model:** sd3.5-medium
- **Encoder:** OpenCLIP-ViT-H/14
- **Negative Prompt:** common-sense, logical
- **Steps:** 20(images per batch)

- **Guidance Scale:** 8.0
- **Resolution:** 512 x 512

### 3.4 VQA dataset Construction

We built a dataset of 400 images and their corresponding questions, aiming to provide a rigorous and discriminative testbed for the evaluation of visual language models. Specifically, each image is paired with a question with two options: one is the common-text generated before replacement in the uncommon-text generation step, and the other is the uncommon-text after replacement. As shown in Figure 3, both options are based on the same image, and the main content of the question remains exactly the same, with the only difference being the text of the options. This design not only helps to clearly distinguish between common sense and uncommon-sense semantic expressions, but also allows us to focus more on the model’s ability to judge subtle semantic differences while eliminating other interference factors.

Which option best describe the input image?



- A. a sheep is chasing a wolf.  
B. a wolf is chasing a sheep.



- A. a grandfather reads storybook for his grandson.  
B. a little boy reads storybook for his grandpa.

Figure 3: examples of a VQA tasks in the UAIT dataset. Each image corresponds to a question and two options. To eliminate the influence of CoT when the model answers, the end of the question contains *”Output only the corresponding letter of the option without providing a reason.”*

During the dataset construction process, we paid special attention to how to guide large

language models to use Chain of Thought (CoT)(16) when reasoning. To achieve this goal, we adopted a prompt engineering strategy to control the output form of the model by designing precise prompts. Specifically, we added specific prompts at the end of each question, requiring the model to only return the final answer without showing its reasoning process. On the one hand, this design avoids exposing the model’s internal logic during the answering process, making the output more concise and clear; on the other hand, it can also objectively evaluate the model’s decision accuracy when it does not rely on chain thinking to explicitly display the reasoning process, thus providing a new evaluation perspective for studying the implicit reasoning ability of the model.

Overall, the construction and question design of this dataset reflects our in-depth exploration of the semantic differences between common sense and counter-common sense in our research, and also ensures the uniformity and effectiveness of the model output during the evaluation process through sophisticated prompt engineering methods. This not only provides high-quality data support for subsequent related experiments, but also lays a solid foundation for exploring the performance differences of large language models under different prompt strategies.

## Chapter 4: Results and Discussion

We designed a new and challenging uncommon action image-text question answering task (UAIT) to systematically evaluate the ability of the most advanced visual language models and models based on contrastive learning strategies to handle complex semantic reasoning problems. The uniqueness of this task lies in the deliberate setting of a series of scenarios in which common sense logic is subverted or even reversed, which not only breaks the conventional logical chain that the model relies on in traditional reasoning tasks, but also forces the model to make new understandings and judgments when faced with typical information combinations.

In this way, we can more intuitively reveal the weak links of existing models beyond conventional semantic processing, and provide practical data support and research directions for future model improvements. The experimental results clearly show that when common sense logic is deliberately reversed, both mainstream visual language models and models based on contrastive learning generally show significant lack of understanding ability. This result reflects that current technology still faces great challenges in breaking through the inherent semantic cognitive mode.



Which option best describes the input images?

A. a suspect is following a policeman  
B. a policeman is following a suspect

The correct description for the image is:  
**B. A policeman is following a suspect.**

In this image, the policeman is walking in front, and the individual in a black shirt appears to be walking behind him.

Figure 4: An example of model answer on the VQA task in UAIT dataset, we use qwen2-vl-instruct as a base model here.

## 4.1 Model selection

In the process of model selection, we selected several mainstream visual language models from the most representative models in the current field: qwen2-VL-Instruct, LLaMA3.2-Vision(17), and LLaVA-1.5. These models have shown high performance in many tasks in the past, so they have become our first choice for visual question answering (VQA) task evaluation, an example can be seen in Figure 4. In addition, in order to further compare and analyze the performance of the model under different training paradigms, we also introduced CLIP and RWKV-CLIP, which are based on contrastive learning. The specific model introduction is shown in Table 4.

Table 4: Information of the models we use in evaluation

model	Training scale	Training method	architecture
Qwen2-VL-Instruct	18 trillion tokens	Pretraining, Instruction Tuning	single-stream
LLaVA1.5	1.2M samples	Feature alignment, Finetuning	single-stream
LLaMA3.2-Vision	60.3M samples	Pretraining, Instruction Tuning	dual-stream
CLIP	400M samples	Contrastive learning, Feature alignment	dual-stream
RWKV-CLIP	15M samples	Contrastive learning, Feature alignment	dual-stream

Since contrastive learning models are not essentially designed to complete prediction tasks similar to visual language models, they have certain limitations when directly performing question answering tasks. For this reason, we use a zero-shot evaluation method to measure the performance of contrastive learning models. Specifically, we removed the question description (*Instruct*) in the VQA dataset during the evaluation process, and directly used the content in *option A* and *option B* as the input of the model. We then selected the most confident answer returned by the model as its prediction, thus achieving a fair comparison method for models without additional fine-tuning.

To further verify human performance on UAIT and provide a reliable comparison benchmark, we invited two participants who had not received professional annotation training to evaluate the entire dataset. The accuracy of the two annotators was statistically processed and the average was taken as the human benchmark on this dataset, providing an intuitive reference for the performance of the model.

## 4.2 Result analysis

The results of VLMs can be found in Table 5. Although Qwen2-VL-Instruct (7B) and LLaMA3.2-Vision (11B) achieved 0.64 and 0.69 accuracy respectively after enabling Chain-of-Thought (CoT) reasoning, showing that the model can perform a certain degree of semantic parsing and logical inference with the assistance of reasoning chain prompts, these two values are still far from the 0.96 accuracy of human annotators. This gap shows

that the current visual language model’s inherent understanding and reasoning ability is still far from human level when dealing with scenes where uncommon sense and common sense are deliberately reversed. In contrast, LLaVA1.5 only achieved an accuracy of 0.36 in the default state, further highlighting that the model is easily limited by the inherent statistical co-occurrence relationship in the training corpus without being optimized specifically for complex semantic problems.

From the comparison of the amount of training data for each model, it can be seen that LLaVA1.5, although only a very small amount of data was used for training, has demonstrated similar capabilities to Qwen2-VL-Instruct and LLaMA3.2-Vision in general question-answering tasks. However, this general question-answering ability obtained based on training with a small amount of data has not been fully demonstrated on UAIT. In other words, when the model faces uncommon scenes, it will appear to be powerless if there is a lack of additional semantic reinforcement prompts or specialized training. For example, common statistical preferences in the training process (“tiger drags rabbit” appears more frequently than “rabbit drags tiger”) may cause the model to misjudge during actual reasoning, resulting in incorrect image descriptions.

To further explore this phenomenon, we divided the UAIT dataset into 70 percent training set and 30 percent test set, and fine-tuned LLaVA1.5 with LoRA. It is exciting that after LoRA fine-tuning, the accuracy of LLaVA1.5 has been significantly improved to 0.79, which demonstrates that even models with smaller parameter scales can significantly enhance their ability to understand complex semantic scenes through customized fine-tuning for targeted tasks. This finding has important implications for applications in low-resource environments: when data and computing power are limited, a refined fine-tuning strategy can effectively make up for the shortcomings of the model in dealing with uncommon-sense problems.

Table 5: Accuracy of vision-language models (VLMs) on the UAIT dataset.

Model	Parameters	Accuracy
CLIP	0.4B	0.49
RWKV-CLIP	0.15B	0.53
Qwen2-VL-Instruct	7B	0.64
LLaVA1.5	7B	0.36
LLaMA3.2-Vision	11B	0.69
human annotation		<b>0.96</b>

Table 6: Accuracy of LLaVA-UAIT.

Model	Parameters	UAIT-Test	Winoground
LLaVA1.5	7B	0.40	0.42
LLaVA-UAIT	7B	<b>0.85</b>	<b>0.48</b>

More importantly, even LLaMA3.2-Vision, which has a much more larger parameter scale,

has not outperformed the fine-tuned LLaVA1.5 without additional fine-tuning. This result reveals an important conclusion that, the current bottleneck of visual language models is not only determined by the size of the model, but also by whether the model has the deep ability to understand the semantic role structure and deal with common sense reversal. In the task of visual action semantic understanding, it is often unreliable to rely solely on the statistical co-occurrence rules in the training data for inference. Only through targeted training and prompt optimization can the model make judgments closer to human levels in complex semantic scenes.

### 4.3 Analysis of comparative learning models

The inference results of the contrastive learning model are shown in Table 7. The accuracy of traditional CLIP in this task is only slightly higher than random guessing, which shows that its discrimination ability in uncommon scenes is very limited. Further experimental results show that the performance of RWKV-CLIP is even lower than that of traditional CLIP. This result intuitively reveals that although RWKV-CLIP may have certain advantages in some open domain image-text matching tasks, its model structure does not show any unique superiority when facing unconventional and logically reversed image-text matching problems.

Table 7: Binary classification accuracy of contrastive learning models on UAIT.

<b>Model</b>	<b>Vision Encoder</b>	<b>Language Model</b>	<b>Binary Accuracy</b>
RWKV-CLIP	Vision Transformer	Transformer	0.49
CLIP	Vision Transformer	RWKV	0.53

There are two key issues hidden behind this result.

First, the image-text alignment model has fundamental flaws in semantic logic modeling. Taking CLIP and its variants as an example, this type of model mainly relies on the similarity between images and texts in the embedding space for judgment, and this method is powerless when dealing with problems involving syntactic role conversion. For example, when the subject and object in a sentence are exchanged (such as agent and patient), although the surface vocabulary of the text has not changed, the semantic meaning has changed fundamentally. Because the model relies too much on the statistical similarity between embedding vectors, it is unable to capture this subtle but crucial logical relationship, which greatly reduces its performance in non-common sense scenarios.

Secondly, the current models' understanding of visual details remains superficial. They often only focus on whether certain key elements (such as "policeman" and "suspect") exist in the image, but cannot accurately parse the dynamic and logical relationships between these elements. For example, in a complex scene, the model may only detect the

presence of "policeman" and "suspect" in the image, but cannot tell whether "policeman is chasing suspect" or "suspect is following policeman". This lack of in-depth understanding of action logic and semantic roles makes the model show obvious limitations when dealing with graphic reasoning tasks involving behaviors and interactive relationships.

Therefore, it can be considered that such models are more suitable for handling static co-occurrence judgment tasks, such as image annotation or image-text retrieval, where we only need to focus on whether elements appear at the same time, without having to reason about the deep logical relationship between elements. On the contrary, for image-text tasks that require fine-grained action logic reasoning, such as the subtle semantic role reversal problem presented in the UAIT dataset, the embedding space of these models obviously lacks sufficient detailed semantic alignment capabilities, resulting in a significant reduction in their matching capabilities. This further emphasizes the necessity of constructing a model structure or training target with logic and role recognition capabilities.

#### 4.4 Discussion on performance between human and models

According to the evaluation results of UAIT by two independent annotators we invited, humans performed significantly better than models in this task, which far exceeded the performance of all existing models in similar tasks. This significant contrast not only proves the excellent ability of humans in understanding and reasoning about unconventional semantic scenes, but also reflects the limitations of current visual language models in complex situational judgments, and further strengthens the following important understandings:

Firstly, although current visual language models have achieved excellent results in standard visual question answering (VQA) tasks and image description generation tasks, they are unable to perform well when faced with problems that require the integration of causal relationships, physical common sense, and semantic role knowledge for situational judgment. In other words, these models often fail to break through the surface phenomenon of statistical co-occurrence when dealing with situations involving deep semantic reasoning and complex logical judgments, resulting in serious deficiencies in uncommon scenarios.

Secondly, even if the Chain-of-Thought strategy is adopted in the model reasoning process, attempting to improve the judgment accuracy by analyzing the problem step by step, the experimental results show that the model still finds it difficult to effectively distinguish the rationality differences caused by semantic reversal. This phenomenon suggests that the reasoning path of the current model relies heavily on the inherent statistical bias in the training corpus, rather than on a deep understanding of semantic connotation and causal logic. Therefore, even with the introduction of the CoT strategy, the model's reasoning

process is still fragile and difficult to make correct judgments when dealing with complex scenarios involving role switching or semantic reversal.

Thirdly, although the fine-tuning method can significantly improve the overall performance of the model to a certain extent, this improvement is often highly dependent on the data construction strategy adopted. In other words, although the effect of fine-tuning training can bring obvious improvements in specific tasks, it cannot fundamentally solve the problem of insufficient generalization ability of the model when facing "beyond common sense" graphic reasoning tasks. The role of fine-tuning methods in compensating for statistical bias is limited, because their training objectives and data distribution are still limited to the inherent laws of training samples, and cannot effectively cover those extremely challenging scenarios that require deep semantic understanding.

In summary, the performance of humans in the UAIT dataset not only provides us with a clear performance comparison benchmark, but also exposes the shortcomings of current visual language models in terms of causal reasoning, physical common sense, and semantic role switching. This result calls for future researchers to make improvements in model design and training methods, so as to achieve cognition and judgment closer to human levels in graphic and text reasoning tasks.

## 4.5 Evaluating and Training configuration

- **Hardware:** 1×A100 80GB GPU
- **Optimization:** AdamW with DeepSpeed-ZoRO3.
- **Learning Rate:** 1e-4
- **Batch Size:** 1 sample (per GPU)
- **gradient accumulation:** 8 steps

## 4.6 Role of UAIT in promoting semantic understanding

The design of UAIT abandons the reliance of traditional VQA tasks on "normal scenes" and for the first time takes "semantic anomaly" as the core focus of evaluation, creatively constructing a set of two-choice questions that are highly symmetrical in structure and have obvious semantic opposition. This new design concept not only subverts the previous evaluation method that relies on the co-occurrence relationship between intuitive images and texts, but also examines the model's reasoning and judgment ability when faced with unconventional semantics and logical reversals. Its main advantages are reflected in the following aspects:

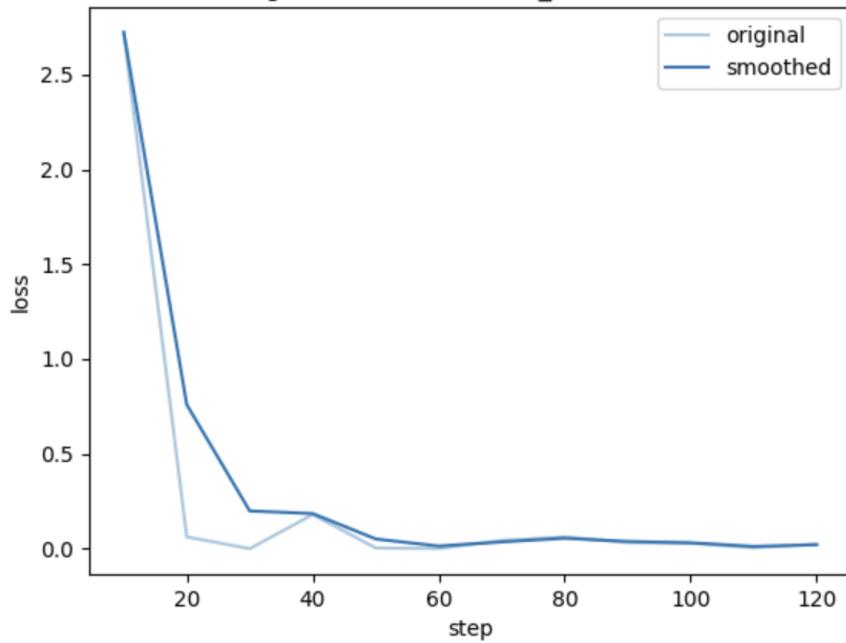


Figure 5: lora finetuning loss of LLaVA-1.5 on UAIT

#### 4.6.1 Stripping away contextual statistical bias

Traditional models often rely on the inherent statistical laws in a large amount of corpus, and habitually select character combinations that appear frequently in training. UAIT effectively strips away common contextual statistical biases by replacing the roles of agent and patient, making it impossible for the model to simply rely on the inherent co-occurrence relationship in the data for judgment. This design forces the model to deeply understand the true semantic relationship between images and texts, rather than making choices based solely on probabilistic preferences.

#### 4.6.2 Higher explainability and diagnosability

Since all questions in UAIT have highly consistent structures and rigorous designs, any model error in answering can be clearly attributed to the failure of semantic reasoning rather than ambiguity in question understanding. This high degree of consistency not only enhances the interpretability of test results, but also provides accurate feedback information for subsequent model diagnosis and optimization, allowing us to more clearly identify the problem.

## 4.7 Implications for further research on multimodal understanding

### 4.7.1 Improving the capabilities on modeling semantic role

The current model mainly focuses on the alignment of visual entity recognition and language description, but lacks systematic modeling of the role reasoning of "agent" and "patient". Future models should integrate stronger syntactic analyzers and semantic role taggers, or introduce structured knowledge graphs to support role recognition and abnormality judgment.

### 4.7.2 Introducing common sense anti-interference mechanism

Many multimodal models still rely on the default assumption of "empirical correctness" when making answer selections, that is, giving priority to options that appear more frequently in the training set. This strategy may be effective in most common situations, but when encountering some situations that are inconsistent with common sense or have abnormal phenomena, the model is prone to misjudgment. The model lacks an effective distinction between "factual correctness" and "empirical correctness", resulting in insufficient judgment when facing abnormal or abnormal situations.

To this end, future research should focus on introducing counter-common sense robustness training mechanisms to enable the model to have:

**The ability to question the bias of the training set:** During the training process, by designing counter-common sense samples and scenarios, the model is prompted to learn to jump out of the limitations of frequency bias and conduct more rigorous fact verification for each input.

**Common sense anti-interference ability:** By constructing a special adversarial training task, the model is forced to identify the irrationality in the face of common sense conflicts and abnormal information, and make judgments that are more in line with factual logic.

Such a mechanism can not only improve the judgment accuracy of the model in complex and abnormal scenarios, but also promote multimodal systems from simply relying on data statistics to a deeper level of common sense understanding and reasoning.

### 4.7.3 Data-driven along with reasoning-driven

The iteration of current multimodal visual language models relies heavily on the expansion of massive data and the capture of statistical laws. However, purely data-driven

methods have certain limitations, especially in tasks that require deep causal reasoning and semantic role understanding. The future development trend should be to combine data-driven with logical reasoning to build a "causal-role" understanding path.

## Chapter 5: Conclusion and Further Work

### 5.1 Conclusion

Our work systematically reveals the core defects of current mainstream visual language models in semantic role understanding, causal reasoning, and abnormal scene judgment for the first time by constructing a picture-text question-answering dataset UAIT with uncommon-sense characteristics.

In the face of semantic role swaps and common sense reversals, the judgment competence of current mainstream visual language models is far below the human level, and they rely heavily on statistical co-occurrence rather than real understanding. By comparing training paradigms, introducing LoRA fine-tuning, and setting human benchmarks, we not only clarified the boundaries of existing models, but also verified the significant gains of fine-tuning on semantically complex tasks.

This research result provides a diagnosable and iterative evaluation framework for future multimodal models to move from "surface matching" to "deep understanding", and points out a breakthrough direction: constructing a new model architecture and anti-bias training mechanism with logical reasoning and role recognition capabilities to promote its evolution in a direction closer to human cognition.

### 5.2 Reflection

Throughout the implementation of this project, I not only mastered the construction and evaluation process of multimodal models, but also systematically exercised technical capabilities such as natural language processing, computer vision, image generation, data annotation and fine-tuning training.

In the process of constructing UAIT, I learned how to design complex language-image pairing tasks based on semantic roles, and to collaboratively generate high-quality counterintuitive samples through large language models and diffusion models. This is a challenging task that I have never been involved in before, which greatly broadened my technical horizons.

At the knowledge level, I have a deeper and more critical understanding of the training logic, reasoning methods and performance boundaries of current mainstream visual language models. I realized that when faced with complex semantic structures or counter-intuitive situations, the model is easily limited by statistical biases in the training corpus, thereby showing irrational or even wrong judgment behaviors. This is not only a technical issue, but also reflects the fundamental challenges of multimodal artificial intelligence

systems in cognitive modeling.

During the project, I also deeply realized the importance of good scientific research practices. From data generation to model testing, every step must adhere to the principles of rigor, transparency and reproducibility to avoid "model cheating" or unconsciously introducing bias. At the same time, I adopted strategies such as structural symmetry design and control variable method, striving to make the evaluation results truly reflect the semantic understanding ability of the model, rather than being induced or affected by prompt words or templates.

From a broader social and ethical perspective, the problems revealed by this project are of practical significance. As multimodal large models are applied to sensitive fields such as education, medical care, and justice, their robustness and credibility in dealing with atypical scenarios become particularly critical. Model misjudgment leads to inefficiency, but is also likely to cause misleading and harm. Therefore, developing AI systems with anomaly detection and semantic discrimination capabilities is not only a technical issue, but also an ethical responsibility.

In addition, during the data construction and model testing phase, I followed the principles of legality and compliance, using domestic and open source large models. The images contained in the data set are all generated by open source models, avoiding the use of any restricted data sources, and ensuring that all generated data do not infringe copyright or privacy.

This project both improved my technical capabilities and theoretical depth, and prompted me to establish a responsible AI development concept in practice. It made me realize that if future AI systems are to truly serve society, they must strike a balance between performance, explainability, and ethical compliance.

### 5.3 Further work

UAIT still has a lot of room for improvement in terms of depth and breadth. Future work will continue to expand and deepen in the following directions:

- **Expanding the scale of UAIT:** Although UAIT can initially reveal the shortcomings of the model in counterintuitive situations, it is still limited in terms of semantic breadth and cultural diversity. In the future, the scale of the dataset will be further expanded through a combination of large-scale automatic generation and manual screening, with the goal of building a medium-to-large benchmark set to improve its statistical stability and challenge.
- **Constructing a multi-dimensional evaluation system:** UAIT currently mainly

uses a two-choice (A/B) question-answering method to evaluate the model's ability to understand semantic role reversal. In the future, it plans to introduce richer evaluation forms, such as open-ended answers, image-text generation matching evaluation, action semantic positioning, etc., to build a more complete evaluation matrix, which not only judges "whether the choice is correct", but also focuses on "why the choice is correct" and "whether it can be explained". In addition, user research will be introduced to compare the similarities and differences between humans and models in cognitive paths and error types, and further improve the diagnostic analysis framework.

- **Introducing more diverse counterintuitive scenarios:** We will explore more types of counter-common sense logic, such as physical counter-intuitive scenarios. These new scenarios will challenge the model's common sense system, causal reasoning, and multimodal fusion competence from different levels, thereby comprehensively promoting the intelligent advancement of VLMs.
- **Integrating multilingual context:** The current research focuses on English scenarios, but the definition of common sense and abnormality itself has cultural differences. In the future, we will explore how to build a multilingual version of UAIT, introduce languages such as Chinese, and conduct comparative studies on the "uncommon sense" judgment ability of cross-cultural models in different cultural backgrounds. This will not only help to test the universality of the model, but also respond to the local adaptation problem in the global deployment of AI.
- **Build training and enhancement mechanisms to improve model capabilities:** Except for evaluation, we will also explore UAIT-based training methods, such as building a counter-common sense corpus enhancement module and introducing role logic adversarial learning strategies to help the model gradually move from "statistical understanding" to "causal reasoning". Especially in scenarios with limited parameter scale or limited resources, this type of targeted training strategy is expected to significantly improve the actual expressiveness and application reliability of the model.

## References

- [1] OpenAI, Achiam J, Adler S, Agarwal S, Ahmad L, Akkaya I, et al.. GPT-4 Technical Report; 2024. Available from: <https://arxiv.org/abs/2303.08774>.
- [2] Liu H, Li C, Wu Q, Lee YJ. Visual Instruction Tuning. In: Oh A, Naumann T, Globerson A, Saenko K, Hardt M, Levine S, editors. Advances in Neural Information Processing Systems. vol. 36. Curran Associates, Inc.; 2023. p. 34892-916. Available from: [https://proceedings.neurips.cc/paper\\_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf](https://proceedings.neurips.cc/paper_files/paper/2023/file/6dcf277ea32ce3288914faf369fe6de0-Paper-Conference.pdf).
- [3] Wang P, Bai S, Tan S, Wang S, Fan Z, Bai J, et al.. Qwen2-VL: Enhancing Vision-Language Model's Perception of the World at Any Resolution; 2024. Available from: <https://arxiv.org/abs/2409.12191>.
- [4] Ronchi MR, Perona P. Describing Common Human Visual Actions in Images; 2015. Available from: <https://arxiv.org/abs/1506.02203>.
- [5] Thrush T, Jiang R, Bartolo M, Singh A, Williams A, Kiela D, et al. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR); 2022. p. 5238-48.
- [6] Rombach R, Blattmann A, Lorenz D, Esser P, Ommer B. High-Resolution Image Synthesis with Latent Diffusion Models; 2022. Available from: <https://arxiv.org/abs/2112.10752>.
- [7] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, et al.. Attention Is All You Need; 2023. Available from: <https://arxiv.org/abs/1706.03762>.
- [8] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, et al.. Learning Transferable Visual Models From Natural Language Supervision; 2021. Available from: <https://arxiv.org/abs/2103.00020>.
- [9] Dosovitskiy A, Beyer L, Kolesnikov A, Weissenborn D, Zhai X, Unterthiner T, et al.. An Image is Worth 16x16 Words: Transformers for Image Recognition at Scale; 2021. Available from: <https://arxiv.org/abs/2010.11929>.
- [10] Sherstinsky A. Fundamentals of Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) network. Physica D: Nonlinear Phenomena. 2020;404:132306. Available from: <https://www.sciencedirect.com/science/article/pii/S0167278919305974>.

- [11] Gu T, Yang K, An X, Feng Z, Liu D, Cai W, et al.. RWKV-CLIP: A Robust Vision-Language Representation Learner; 2024. Available from: <https://arxiv.org/abs/2406.06973>.
- [12] Hu EJ, Shen Y, Wallis P, Allen-Zhu Z, Li Y, Wang S, et al.. LoRA: Low-Rank Adaptation of Large Language Models; 2021. Available from: <https://arxiv.org/abs/2106.09685>.
- [13] Palmer M, Bonial C, Hwang J. 315VerbNet: Capturing English Verb Behavior, Meaning, and Usage. In: The Oxford Handbook of Cognitive Science. Oxford University Press; 2017. Available from: <https://doi.org/10.1093/oxfordhb/9780199842193.013.15>.
- [14] Yang A, Yang B, Hui B, Zheng B, Yu B, Zhou C, et al.. Qwen2 Technical Report; 2024. Available from: <https://arxiv.org/abs/2407.10671>.
- [15] Brown TB, Mann B, Ryder N, Subbiah M, Kaplan J, Dhariwal P, et al.. Language Models are Few-Shot Learners; 2020. Available from: <https://arxiv.org/abs/2005.14165>.
- [16] Wei J, Wang X, Schuurmans D, Bosma M, Ichter B, Xia F, et al.. Chain-of-Thought Prompting Elicits Reasoning in Large Language Models; 2023. Available from: <https://arxiv.org/abs/2201.11903>.
- [17] Gao P, Han J, Zhang R, Lin Z, Geng S, Zhou A, et al.. LLaMA-Adapter V2: Parameter-Efficient Visual Instruction Model; 2023. Available from: <https://arxiv.org/abs/2304.15010>.