Benchmarking Multimodal Idiomaticity: Tasks and Methods for Idiomatic Language Understanding in Text and Images

Anonymous ACL submission

Abstract

In this paper, we present a dataset containing images and texts representing potentially idiomatic expressions in two languages, English and Portuguese. The expressions were selected for their potential of ambiguity between a literal and an idiomatic sense, and they are represented as static images, or as image sequences, to capture the more abstract cases or temporally dependent cases. To investigate how well models handle idiomatic expressions and integrate cues from different modalities (textual and visual/visual-temporal data), we propose two tasks to examine how mono and multimodal representations perform: multiple choice image selection and next image prediction task. Using a new metric that we propose for graded relevance, Normalized Discounted Cumulative Gain, the results obtained by representative models indicate that multimodal generative models, using our framework, outperform traditional vision-and-language models in comprehending idiomatic expressions by effectively integrating visual and textual information.

1 Introduction

The ability to effectively represent non-compositional language is a critical challenge in natural language processing (NLP), as misinterpretations can propagate through downstream applications and lead to erroneous outputs (Yazdani et al., 2015). Idiomatic expressions (IE), a prime example of non-compositional language, pose unique difficulties due to their meanings often diverging significantly from their literal interpretations (He et al., 2024b). While recent advances in language modeling have made strides in capturing such phenomena in text (Zeng and Bhat, 2022; Zeng et al., 2023; He et al., 2024a), the evaluation of multimodal representations of idiomaticity remains underexplored. Research on figurative language frequently centers solely on text, even though there may be complementary potentially disambiguating clues in different modalities, such as when combining text and images. This is particularly evident in social media, advertising, and news contexts (Yosef et al., 2023).

To bridge this gap, we introduce two novel benchmark tasks that integrate textual and visual modalities to assess the semantic comprehension of idiomatic expressions: (A) **multiple image choice**, selecting the image that best represents the intended meaning of an idiomatic expression within a sentence, and (B) **next image prediction**, selecting the most appropriate image to complete a sequence of three images, visually representing temporal aspects of the intended meaning of these expressions. The dataset includes context sentences paired with both individual images and image sequences, enabling a comprehensive evaluation of semantic alignment in static and temporal contexts.

Building on previous work (Tayyar Madabushi et al., 2022), which evaluated idiomaticity solely through textual analysis, this task broadens the scope to assess how well models can integrate linguistic and visual information in representing non-compositional meanings. Unlike prior approaches that relied on images from existing databases (Yosef et al., 2023), our task utilizes newly generated images created using large language-vision models in collaboration with human experts. This ensures that the visual content is specifically tailored to accurately represent idiomatic expressions, closely aligning with the intended linguistic context. To allow for a more nuanced evaluation of the ranking of images or captions based on their relevance to the idiomatic meaning, we propose the use of the Normalized Discounted Cumulative Gain (NDCG) metric, adapted from information retrieval (Järvelin and Kekäläinen, 2002). Unlike traditional binary correctness metrics, NDCG captures graded relevance levels and emphasizes the importance of correct ranking order, making it par-

ticularly suitable for assessing nuanced semantic alignment in multimodal tasks.

The increasing reliance on vision-and-language models for text-image matching reflects a growing trend toward specialization in multimodal AI applications. These models, often trained with objectives such as contrastive learning and joint embedding techniques, excel in tasks like retrieval, ranking, and search by achieving precise and efficient text-image alignment (Hessel et al., 2021; Lee et al., 2024).

However, when it comes to understanding idiomatic expressions, multimodal generative models often outperform traditional vision-and-language models. This advantage stems from their extensive linguistic training and contextual reasoning capabilities (Sun et al., 2024). Trained on diverse and expansive text corpora, these models are adept at interpreting idiomatic expressions, slang, and figurative language. They integrate visual and linguistic semantics, enabling them to determine whether an expression is used literally or figuratively based on contextual cues from both modalities.

We investigate the performance of representative models on these tasks and propose using VQAScore (Lin et al., 2025), a metric for evaluating image-text alignment in visual-question-answering (VQA) models. VQAScore works by posing straightforward queries like "Does this figure show $\{text\}$?" and determining the probability of a "Yes" response, offering a more context-sensitive assessment of alignment. This approach achieves state-of-the-art performance across various benchmarks by utilizing off-the-shelf multimodal generative models.

In this work, we make the following key contributions. First, we propose novel benchmark tasks for multimodal idiomaticity, integrating textual and visual modalities to evaluate the comprehension of idiomatic expressions. This includes a tailored dataset of expert-curated images generated using large language-vision models, with humans in the loop, ensuring precise alignment with idiomatic meanings in two tasks (**multiple image choice** and **next image prediction**). Second, we introduce an adapted NDCG metric for graded relevance assessment of semantic alignment. Finally, we leverage the VQAScore to build baseline methods for the proposed tasks, setting a foundation for future advancements in multimodal idiomaticity research. Understanding idioms is crucial for precise communication and applications such as sentiment analysis, machine translation and natural language understanding. Exploring ways to improve models' ability to interpret idiomatic expressions can enhance the performance of these applications.

2 Related Work

Idioms are believed to be conceptual products and humans understand their meaning from interactions with the real world involving multiple senses (Lakoff and Johnson, 1980; Benczes, 2002). However, investigations about idiomatic understanding by models have mainly concentrated on one modality (textual stimuli), in tasks ranging from evaluation of noun compound paraphrases (Hendrickx et al., 2013), to noun compound interpretation (Butnariu et al., 2009) and compositional models (Marelli et al., 2014), with recent efforts focusing on idiomaticity detection and representation (Madabushi et al., 2022; Garcia et al., 2021b; He et al., 2024b). Indeed recent related datasets for the evaluation of idiomatic and figurative language (e.g., MAGPIE (Haagsma et al., 2020), NCTTI (Garcia et al., 2021a)) concentrate on text, with exceptions like FLUTE (Chakrabarty et al., 2022a)) going beyond text and also includes images, in this case static images representing figurative usages. When models are evaluated against human performance, for their understanding of idiomatic expressions in both textual data (Tayyar Madabushi et al., 2021; Chakrabarty et al., 2022b; He et al., 2024b) as well as in multimodal settings (Yosef et al., 2023), they are still found to lag behind. One possibility is that for potentially idiomatic expressions, which can take on either an idiomatic or a more literal sense[1], contextual cues not only from texts but also from other modalities may be required for succesfully determining the target sense. In this case, the relevance and impact of the contribution of the different modalities involved need to be investigated. Additionally, model performance on these tasks may also be confounded by artifacts present in datasets (Boisson et al., 2023), or in characteristics of the spaces defined by the models (He et al., 2024b) which may lead to an appearance of better performance at idiomaticity detection tasks but without necessarily developing high-quality representations that accurately capture the semantics of idiomatic expressions. This work, building on insights from the text-only task

[1]E.g. *gold mine* in its literal sense, or as the idiomatic *profitable business*.

(Madabushi et al., 2022), explores the idiomatic comprehension ability of multimodal models. In particular, we focus on models that incorporate visual and textual information and how accurately they capture potentially idiomatic expressions and whether multiple modalities can improve these representations. The proposed dataset seeks to address these questions by providing potentially idiomatic expressions along with sentences and images representative of both the literal and idiomatic sense. Moreover, we propose two subtasks that allow the examination of both senses that can be accurately represented by static images and of more abstract and temporally senses that require more. We hope to address these shortcomings by moving away from binary classification and by introducing representations of meaning using visual and visual-temporal modalities.

Using internet-sourced images for model training brings significant challenges (e.g. copyright infringement, privacy violations, and unintended biases), with datasets like LAION-400M often containing explicit content, harmful stereotypes, and unbalanced representation, raising both ethical and legal concerns (Birhane et al., 2021; Crawford and Paglen, 2021). An alternative is to use a generative approach combining a Large Language Model (LLMs) with a diffusion model. By taking advantage of fine-tuned prompts, it may potentially produce balanced and diverse datasets across demographics and scenarios, reducing the risk of bias. This approach, which we adopt in this paper, is also highly scalable and efficient, enabling rapid and cost-effective generation of high-quality, domain-specific images (Resnik and Hosseini, 2024).

3 Task

To combine an LLM with a diffusion model we used Midjourney[2] given the fine-grained human control needed for image generation to produce high-quality, domain-specific images tailored to the task, guided by human expert supervision to ensure alignment with literal and idiomatic meanings.

3.1 Task A: Multiple Image Choice

Given a context sentence containing a potentially idiomatic nominal compound (NC) and a set of five images, the task is to rank the images based on how accurately they depict the meaning of the NC used in that sentence. A variation of task also allows for monomodal settings, where given a sentence and five text captions (each describing the content of one of the images, as described in Section 4.3) the goal is to rank the captions on how they capture the meaning of the NC. Figure 1 provides an example of the Subtask A data for the expression *bad apple*.

The idiomatic expressions used in the test set will be completely different from those provided in the training data so as to ensure that models learn the ability to generalise as opposed to "memorising" idiomatic phrases, further emphasised through the zero-shot component. The English dataset for Subtask A includes 70 training items (350 image-caption pairs), 15 development items, and 15 test items, while the Portuguese dataset comprises 32 training items (160 image-caption pairs), 10 development items, and 10 test items.

3.2 Subtask B: Image Sequences (or Next Image Prediction)

Capturing the idiomatic meaning of an MWE in a single image is not necessarily straightforward. While one can envisage a literal *kangaroo court*, a good representation of its idiomatic sense would need to incorporate elements (spontaneity, haste, a potentially predetermined conclusion) which are less concrete than a marsupial wielding a gavel.

In order to better represent the abstract meaning of our target expressions, we propose to generate sequences of 3 images akin to a comic strip, allowing for the depiction of changes in state, mood or relationship between elements over time.

In Subtask B, a target expression and an image sequence with the final image removed will be provided, along with a set of candidate images. The task is to select the most suitable image to complete the sequence while also determining whether the depicted sense of the nominal compound (NC) is idiomatic or literal. Examples are shown in Figure 2. The sense of the compound (literal or idiomatic) depicted in the sequence will also be indicated.

In order to minimise the risk of non-semantic clues being introduced, the images will adopt a consistent style across the Subtask B dataset. As with Subtask A, we will also offer two settings for Subtask B, with descriptive text replacing the images in the 'caption' setting. In the Subtask B dataset, the English set includes 20 examples for training, 5 for development, and 5 for testing, while the Portuguese set includes 15 examples for training, 5 for development, and 5 for testing.

[2]https://www.midjourney.com/

(a) The image depicts three children standing in front of a gray, textured wall...

(b) The image depicts a cartoon-style illustration of a young boy standing at a table...

(c) The image depicts a halved peach with a detailed and realistic appearance...

(d) The image depicts an orange-colored apple that appears to be decomposing or decaying...

(e) The image depicts a rustic, burlap sack filled with several bright orange apples...

Figure 1: Subtask A data example for *bad apple*. Images were generated using Midjourney, with the style guidance and the prompt completions shown. Captions are displayed partially. The complete example can be found in the Appendix.

First two of the Sequence:

(a) The image shows a classroom scene with five animated characters, likely children, sitting at desks...

(b) The image shows an animated character, a young boy with red hair, sitting at a desk with a laptop...

Candidates:

(c) The image shows a group of animated characters, likely children, standing in front of a television screen that displays various cartoon animals...

(d) The image shows an animated character, a young man with brown hair, standing next to a trash can...

(e) The image shows a cartoon illustration of a young girl with red hair, wearing a white sweater with a red bow and a plaid skirt...

(f) The image shows a group of animated characters that appear to be in a state of distress or chaos...

Figure 2: Subtask B data example for *bad apple*. Images (a) and (b) form the initial part of the sequence, while images (c) through (f) serve as candidates.

4 Data and Resources

Our project will use a dataset which expands on the SemEval-2022 Task 2 dataset (Tayyar Madabushi et al., 2022). Data will be licensed under Creative Commons Attribution-ShareAlike 4.0.

4.1 Subtask A Data

For each idiom a set of 5 different images will be generated with a fixed style prompt within each set to ensure consistency. The images generated for each expression will use the following prompts:

- A paraphrase or representation that captures the idiomatic meaning of the NC.
- A synonym for the literal meaning of the NC.
- Something related to the idiomatic meaning, but not synonymous.
- Something related to the literal meaning, but not synonymous.
- A 'distractor' belongs to the same category as the compound (e.g. an object or activity) but is unrelated to both the literal and idiomatic meanings.

Figure 1 shows an example of the Subtask A data for the expression *bad apple*. For a sentence in which *bad apple* is used idiomatically ("However, it will not work unless every single person does it, because one bad apple ruins the whole barrel."), the expectation is that the images will be ordered as shown in Figure 1, with the idiomatic synonym of corrupting influence ranked as most similar to the in-context sense.

4.2 Subtask B Data

A sequence of images are generated for each NC: one sequence representing the literal or the idiomatic meaning (Figure 2). Each image in a sequence is generated individually using manually crafted prompts (Details are shown in Appendix A) by Midjourney, inspired by the work of Chakrabarty et al. (2023) on visual metaphors, and styled consistently for uniformity across the data.

4.3 Generating Captions by Prompting LLMs

In this study, we utilize the *LLaVA-HF/v1.6-mistral-7b-hf*[3] (LLaVA) model, a vision-language large language model specifically designed for tasks requiring multimodal reasoning (Liu et al., 2024). LLaVA integrates a vision encoder to extract semantic features from images and a large language model to process these features and generate text. By employing the prompt "*What is shown in this image?*", the model generates captions that describe the content of the input images. The workflow ensures that the visual and textual components of the model work in harmony to produce accurate and contextually relevant descriptions. To ensure the quality of the generated captions, all outputs are reviewed and verified by human evaluators.

4.4 Evaluation

As mentioned in the last section, the dataset produced to be used in our experiments was generated with LLMs under human expert verification, in order to guarantee high-quality, accurate, and relevant content. This rigorous human involvement during dataset creation serves as a benchmark of reliability, eliminating the immediate need for additional human evaluation. Our focus is then to use this human-verified dataset to objectively evaluate model performance, with future work potentially incorporating further human assessments to extend our findings.

4.4.1 Subtask A

Performance for Subtask A will be assessed with two key metrics: a) Top Image Accuracy, that measures the correct identification of the most representative image and b) the Normalized Discounted Cumulative Gain (NDCG) (Järvelin and Kekäläinen, 2002) an established information retrieval metric that not only captures the fraction of retrieved relevant information but also takes into account their correct ordering. The *Discounted Cumulative Gain (DCG)* is defined as

$$\mathrm{DCG}_n = \sum_{i=1}^{n} \frac{rel_i}{\log_2(i+1)}, \tag{1}$$

where rel_i is the relevance score of the i-th item, and n is the number of items considered. Its normalized version, which ranges between 0 and 1, is defined as

$$\mathrm{NDCG}_n = \frac{\mathrm{DCG}_n}{\mathrm{IDCG}_n}. \tag{2}$$

[3]https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

where IDCG is the ideal (maximum) possible DCG for the same set of items ranked in the optimal order (highest relevance first). A value of 1 corresponds to a perfect ranking, while lower values reflect less optimal rankings. NDCG is particularly suitable for this task because it provides a nuanced, rank-based measure that captures varying degrees of relevance among multiple candidate images, encouraging models to produce rankings that align more closely with human understanding of idiomatic meaning.

4.4.2 Subtask B

This subtask assesses the model's ability to complete a sequence of images that narratively represent an idiomatic expression, along with distinguishing between idiomatic and literal meanings. Evaluation metrics will be a) Completion Accuracy, that measures the correctness of the selected image to complete the narrative and b) Labeling F1 Score that measures the effectiveness in identifying idiomatic versus literal expressions.

5 Methods

Recent advancements in multimodal generative models, such as LLaVA (Liu et al., 2024), InstructBLIP (Dai et al., 2024), OpenFlamingo (Awadalla et al., 2023), and MIMIC-IT (Li et al., 2023), have shown significant potential in integrating vision and language for diverse tasks. While understanding idiomatic expressions (IEs) demands semantic reasoning, contextual comprehension, and commonsense knowledge (Phelps et al., 2024), these models still struggle to fully capture task-specific nuances. In particular, challenges persist when reasoning across multiple objects, attributes, and relations (Kamath et al., 2023; Lin et al., 2024; Lu et al., 2024; Wang et al., 2023; Yuksekgonul et al., 2022).

To address these challenges, we propose a novel adaptation of Visual Question Answering (VQA) methods for idiomatic image-text alignment tasks. Inspired by recent advances in text-image alignment (Yarom et al., 2024) and text-to-visual generation evaluation (Lin et al., 2025), we introduce a zero-shot approach for automatically evaluating idiomatic image-text alignment using VQA. This approach avoids the need for task-specific fine-tuning, making it scalable and efficient while ensuring high accuracy in aligning idiomatic meanings across modalities.

6 Experiments

To validate the effectiveness and generalizability of our proposed methods, we conducted extensive experiments on the benchmark dataset designed for multimodal idiomaticity, as well as on an external dataset, IRFL: Image Recognition of Figurative Language (Yosef et al., 2023). The experiments were designed to assess both subtasks independently and to explore the ability of our methods to handle figurative and literal interpretations across different scenarios.

6.1 Subtask A

Subtask A evaluates a model's ability to interpret noun compounds (NCs) across a spectrum of literal to figurative meanings within various contexts. The task requires selecting the image that best represents the contextual meaning of the NC in a given sentence. To achieve this, we leverage the VQAScore framework to compute the likelihood that a model assigns a "Yes" response to a dynamically generated query associated with the NC and each candidate image. This likelihood is expressed as:

$$P(\text{“Yes”} \mid \text{Image, Question}) \quad (3)$$

The query is dynamically formulated to probe whether the NC meaning aligns with the image content. Specifically, we use the question: "*Does this figure show the meaning of* $< compound >$ *in the sentence:* $< context\_sentence >$*? Please answer yes or no.*" Among the candidate images $[P1, P2, P3, P4, P5]$, the one with the highest likelihood is selected as the best match. This approach evaluates both the contextual understanding of the NC and the model's ability to interpret semantically relevant information from images. Our method is illustrated in Figure 3b, which adapts the VQAScore framework for multimodal understanding. We use the *clip-flant5-xl*(CFT5) (Lin et al., 2025) model within this framework to achieve the best results. In contrast, Figure 3a depicts the traditional approach for computing matching scores.

For Subtask A, we also implemented methods based on **CLIP-based models** (Figure 3a) and **text-only models**. These models measure the semantic similarity between the query text (an NC with its context sentence) and the candidate images/captions. The CLIP-based models utilize both visual and textual information, while the text-only models rely solely on the captions associated with the candidate images. This comparison helps evaluate the relative contributions of multimodal versus text-only approaches in selecting the image that best matches the meaning of the query text. The results of these experiments are summarized in Table 1.

CLIP-based Multimodal Models These models process and align textual and visual inputs, making them suitable benchmarks for evaluating multimodal understanding:

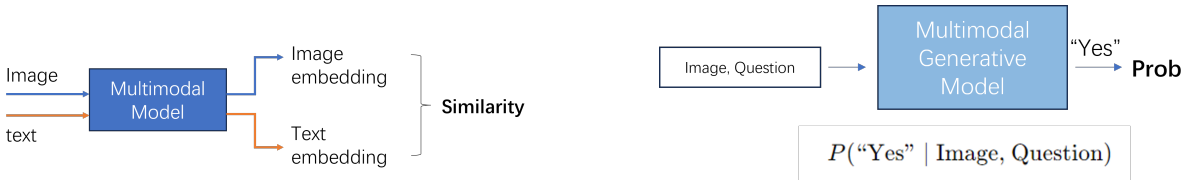
- **openai/clip-vit-large-patch14**[4] (CLIP14): The original CLIP model from OpenAI, built on a ViT-Large architecture with Patch 14, widely used for image-text matching tasks.
- **zer0int/CLIP-zer0int**[5] (zer0int): A variant of CLIP leveraging the ViT-Large architecture with Patch 14, optimized for general-purpose multimodal tasks.
- **UCSC-VLAA/ViT-L-16-HTxt-Recap-CLIP**[6] (UCSC16): A CLIP-based model enhanced with hierarchical text representations (Li et al., 2024), which explores the impact of recaptioning billions of web images.

Text-Only Models These models focus solely on textual inputs, providing a baseline for evaluating semantic understanding and text generation:

- **BAAI/bge-reranker-large**[7] (BAAI-large): A large language model fine-tuned for reranking tasks, designed to optimize semantic alignment in textual data (Xiao et al., 2024).
- **BAAI/bge-reranker-v2-m3**[8] (BAAI-m3): An updated version of the reranker model, incorporating advanced training techniques for improved performance in multilingual and semantic reranking tasks (Chen et al., 2024).

Analysis The performance of the models, as shown in Table 1, highlights distinct differences between text-only and multimodal approaches across both English and Portuguese datasets. On the English dataset, the multimodal model UCSC16

[4]https://huggingface.co/openai/clip-vit-large-patch14
[5]https://huggingface.co/zer0int/CLIP-zer0int
[6]https://huggingface.co/UCSC-VLAA/ViT-L-16-HTxt-Recap-CLIP
[7]https://huggingface.co/BAAI/bge-reranker-large
[8]https://huggingface.co/BAAI/bge-reranker-v2-m3



(a) A multimodal text and image prediction framework.

(b) Our proposed approach.

Figure 3: (a) A multimodal model processes textual and visual inputs to produce predictions. (b) Our approach computes $P(\text{“Yes”} \mid \text{Image, Question})$ using a multimodal generative model.

achieves the highest accuracy (0.47) among comparison models, demonstrating its capability to align visual and textual information effectively. However, its NDCG score (0.76) is lower than zer0int (0.89) and CLIP14 (0.88), indicating that while UCSC16 excels in accuracy, it is less consistent in ranking.

In the Portuguese dataset, UCSC16 again leads among comparison models with an accuracy of 0.50 and an NDCG score of 0.94, showcasing its ability to generalize across languages. CLIP14 and zer0int follow with accuracies of 0.40 and 0.30, respectively, and NDCG scores of 0.91 and 0.90. These results reinforce the general effectiveness of CLIP-based models for multimodal tasks, though UCSC16 shows superior contextual understanding.

For text-only models, BAAI-m3 achieves the highest accuracy (0.50) on the Portuguese dataset, matching UCSC16 but falling behind in NDCG (0.91). On the English dataset, BAAI-large performs best among text-only models with an accuracy of 0.33 and an NDCG of 0.89. These results indicate that text-only models can perform well in ranking tasks but are generally less effective at leveraging contextual information for accurate predictions.

The proposed VQAScore framework outperforms all other models on both datasets. In the English dataset, CFT5 achieves the highest accuracy (0.80) and NDCG (0.95), demonstrating its strong capability to integrate visual and textual information. Similarly, in the Portuguese dataset, LLaVA1.5 achieves the top performance with an accuracy of 0.80 and an NDCG of 0.95, closely followed by CFT5 (0.50 accuracy, 0.92 NDCG). These results validate the generalizability and effectiveness of the VQAScore framework across different languages and contexts, significantly outperforming both multimodal and text-only baselines.

6.2 Subtask B

Subtask B evaluates a model's ability to complete an image sequence by selecting the most suitable image to fill the gap. Each sequence consists of three images, with the final image removed, and a set of candidate images is provided. The goal is to select the candidate image that best aligns with the context established by the first two images in the sequence. This task tests the model's ability to maintain visual and semantic consistency across a narrative while discerning idiomatic or literal interpretations.

The base method for Subtask B builds upon the approach used in Subtask A. Specifically, we adapt the VQAScore framework to compute the likelihood. To achieve this, each candidate image is combined with the first two images in the sequence to form a single long image, which is then used to calculate the likelihood of Equation (3). By treating the combined sequence as a unified input, the model evaluates how well each candidate aligns with the visual narrative established by the preceding images. The candidate image with the highest likelihood is selected as the the best fit for completing the sequence.

Analysis of Results The performance of the models for Subtask B, as presented in Table 2, demonstrates varying capabilities in handling sequence completion tasks. Both InstructBLIP-FlanT5-XL (Dai et al., 2024) and GPT-4o (Achiam et al., 2023) achieve the highest accuracy (0.60) and F1 score (0.43), showcasing their strong ability to align candidate images with the visual context established by the preceding sequence. This suggests that these models are equally effective at maintaining narrative coherence and discerning idiomatic or literal interpretations within image sequences.

In contrast, CLIP-FlanT5-XXL shows significantly lower performance, with an accuracy of 0.20 and an F1 score of 0.11. This indicates substantial limitations in its ability to process and align

visual and textual information for this task. These results highlight the relative strength of multimodal models like InstructBLIP-FlanT5-XL and GPT-4o, while also emphasizing the challenges faced by other models in fully capturing the nuances of sequence-based reasoning. Further improvements in contextual understanding and reasoning are necessary to achieve consistent performance across diverse models.

Model Name	C-only	Accuracy	NDCG
English Dataset			
zer0int	-	0.40	0.89
UCSC16	-	0.47	0.76
CLIP14	-	0.40	0.88
BAAI-large	**y**	0.33	0.89
BAAI-m3	**y**	0.27	0.87
Proposed Method			
CFT5	-	0.80	0.95
Portugues Dataset			
zer0int	-	0.30	0.90
UCSC16	-	0.50	0.94
CLIP14	-	0.40	0.91
BAAI-large	**y**	0.20	0.91
BAAI-m3	**y**	0.50	0.91
Proposed Method			
CFT5	-	0.50	0.92
LLaVA1.5	-	0.80	0.95

Table 1: Performance of various models on Subtask A test datasets in English and Portuguese, showing top image accuracy and NDCG. **C-Only** denotes caption-only models, with sections for each language and the proposed VQAScore framework highlighted.

Model Name	Accuracy	F1
instructblip-flant5-xl	0.60	0.43
clip-flant5-xxl	0.20	0.11
GPT4o	0.60	0.43

Table 2: Performance of different foundation models using the VQAScore framework for on the test dataset for Subtask B. The table reports Accuracy and F1 scores to evaluate the models' ability to complete image sequences by selecting the most suitable candidate image.

6.2.1 Other datasets

To further evaluate the effectiveness of our method beyond the proposed benchmark, we tested it on the IRFL dataset (Yosef et al., 2023). The IRFL dataset is specifically designed to benchmark multimodal understanding of figurative language, including idioms, metaphors, and similes, by pairing textual descriptions with corresponding figurative and literal images. For this evaluation, we focused solely on figurative idioms. This setup provides a targeted platform for evaluating the generalizability of our approach in handling complex, non-literal expressions. The results from this evaluation are shown in Table 3, highlighting both the strengths and limitations of our method compared to existing zero-shot models and human performance.

Categories	Figurative idioms
Humans	97
CLIP-VIT-L/14	17
CLIP-VIT-B/32	16
CLIP-RN50	14
CLIP-RN50x64	22
BLIP	18
BLIP2	19
CoCa ViT-L-14	17
VQAscore(CFT5)	**35**

Table 3: Zero-shot models' performance on the IRFL "mixed" multimodal figurative language detection task for figurative idioms. Numbers represent the percentage of instances correctly annotated. Models fail to reach human-level performance.

The results in Table 3 demonstrate the challenges models face in detecting figurative idioms. Human performance is the highest at 97%, highlighting the complexity of the task. Our proposed VQAscore (CFT5) achieves the best result among zero-shot models with 35% accuracy, significantly outperforming other approaches. This result emphasizes its ability to better align visual and textual information for figurative idiom detection.

The comparison results for models such as CLIP-VIT-L/14 (17%), BLIP2 (19%) and Human evaluations (97%) are taken directly from the IRFL paper (Yosef et al., 2023). These models exhibit limited capabilities in handling figurative language, with marginal differences across variants. Despite VQAscore's relative success, the gap to human performance underscores the need for further advancements in multimodal figurative language understanding.

7 Conclusion

This work introduces benchmarks for multimodal idiomaticity understanding, focusing on static image ranking and image sequence prediction. Our results highlight progress in integrating vision-language models with the novel adapted metric NDCG, though significant gaps remain compared to human performance. These findings lay a foundation for improving LLM's handling of idiomatic expressions across languages and modalities.

8 Limitations

While our work introduces novel benchmark tasks and datasets for multimodal idiomaticity, several limitations should be acknowledged:

Size and Diversity The datasets for both Subtask A and Subtask B are small, which may hinder robust generalization, particularly for idiomatic expressions that are highly context-dependent. The dataset may not fully capture the broad spectrum of idiomatic language, especially subtle or culturally specific expressions.

Task Complexity The next image prediction task assumes that models can discern temporal and abstract relationships effectively. However, existing models struggle with these complexities. Our proposed methods for Subtask B does not explicitly model temporal or abstract relationships, potentially oversimplifying the task.

Limited Use of Training Data Our proposed methods do not utilize the provided training set for fine-tuning or supervised learning, instead relying solely on pre-trained models. This may limit their ability to adapt to task-specific nuances and fully benefit from the curated dataset.

Acknowledgments

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al. 2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Réka Benczes. 2002. The semantics of idioms: a cognitive linguistic approach. *The Even Yearbook*, 5:17–30.

Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963*.

Joanne Boisson, Luis Espinosa-Anke, and Jose Camacho-Collados. 2023. Construction Artifacts in Metaphor Identification Datasets. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6581–6590, Singapore. Association for Computational Linguistics.

Cristina Butnariu, Su Nam Kim, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2009. SemEval-2010 task 9: The interpretation of noun compounds using paraphrasing verbs and prepositions. In *Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-2009)*, pages 100–105, Boulder, Colorado. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022a. FLUTE: Figurative Language Understanding through Textual Explanations. *arXiv preprint*. ArXiv:2205.12404 [cs].

Tuhin Chakrabarty, Arkadiy Saakyan, Debanjan Ghosh, and Smaranda Muresan. 2022b. FLUTE: Figurative language understanding through textual explanations. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 7139–7159, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Tuhin Chakrabarty, Arkadiy Saakyan, Olivia Winn, Artemis Panagopoulou, Yue Yang, Marianna Apidianaki, and Smaranda Muresan. 2023. I spy a metaphor: Large language models and diffusion models co-create visual metaphors. *Preprint*, arXiv:2305.14724.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024. M3-embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 2318–2335, Bangkok, Thailand. Association for Computational Linguistics.

Kate Crawford and Trevor Paglen. 2021. Excavating ai: The politics of images in machine learning training sets. *Ai & Society*, 36(4):1105–1116.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2024. Instructblip: towards general-purpose vision-language models with instruction tuning. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021a. Assessing the representations of idiomaticity in vector models with a noun compound dataset labeled at type and token levels. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2730–2741, Online. Association for Computational Linguistics.

Marcos Garcia, Tiago Kramer Vieira, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2021b. Probing for idiomaticity in vector space models. In

Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pages 3551–3564, Online. Association for Computational Linguistics.

Hessel Haagsma, Johan Bos, and Malvina Nissim. 2020. MAGPIE: A large corpus of potentially idiomatic expressions. In *Proceedings of the 12th language resources and evaluation conference*, pages 279–287, Marseille, France. European Language Resources Association.

Wei He, Marco Idiart, Carolina Scarton, and Aline Villavicencio. 2024a. Enhancing idiomatic representation in multiple languages via an adaptive contrastive triplet loss. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12473–12485, Bangkok, Thailand. Association for Computational Linguistics.

Wei He, Tiago Kramer Vieira, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2024b. Investigating idiomaticity in word representations. *Computational Linguistics*, pages 1–48.

Iris Hendrickx, Zornitsa Kozareva, Preslav Nakov, Diarmuid Ó Séaghdha, Stan Szpakowicz, and Tony Veale. 2013. SemEval-2013 task 4: Free paraphrases of noun compounds. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 138–143, Atlanta, Georgia, USA. Association for Computational Linguistics.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 7514–7528, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated gain-based evaluation of ir techniques. *ACM Transactions on Information Systems (TOIS)*, 20(4):422–446.

Amita Kamath, Jack Hessel, and Kai-Wei Chang. 2023. Text encoders bottleneck compositionality in contrastive vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4933–4944, Singapore. Association for Computational Linguistics.

George Lakoff and Mark Johnson. 1980. The metaphorical structure of the human conceptual system. *Cognitive science*, 4(2):195–208.

Yebin Lee, Imseong Park, and Myungjoo Kang. 2024. FLEUR: An explainable reference-free evaluation metric for image captioning using a large multimodal model. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3732–3746, Bangkok, Thailand. Association for Computational Linguistics.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2023. Mimic-it: Multi-modal in-context instruction tuning. *arXiv preprint arXiv:2306.05425*.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. 2024. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*.

Zhiqiu Lin, Xinyue Chen, Deepak Pathak, Pengchuan Zhang, and Deva Ramanan. 2024. Revisiting the role of language priors in vision-language models. In *International Conference on Machine Learning*. PMLR.

Zhiqiu Lin, Deepak Pathak, Baiqi Li, Jiayao Li, Xide Xia, Graham Neubig, Pengchuan Zhang, and Deva Ramanan. 2025. Evaluating text-to-visual generation with image-to-text generation. In *European Conference on Computer Vision*, pages 366–384. Springer.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Yujie Lu, Xianjun Yang, Xiujun Li, Xin Eric Wang, and William Yang Wang. 2024. Llmscore: Unveiling the power of large language models in text-to-image synthesis evaluation. *Advances in Neural Information Processing Systems*, 36.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 Task 2: Multilingual Idiomaticity Detection and Sentence Embedding. *arXiv preprint*. ArXiv:2204.10050 [cs].

Marco Marelli, Luisa Bentivogli, Marco Baroni, Raffaella Bernardi, Stefano Menini, and Roberto Zamparelli. 2014. SemEval-2014 task 1: Evaluation of compositional distributional semantic models on full sentences through semantic relatedness and textual entailment. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 1–8, Dublin, Ireland. Association for Computational Linguistics.

Dylan Phelps, Thomas M. R. Pickard, Maggie Mi, Edward Gow-Smith, and Aline Villavicencio. 2024. Sign of the times: Evaluating the use of large language models for idiomaticity detection. In *Proceedings of the Joint Workshop on Multiword Expressions and Universal Dependencies (MWE-UD) @ LREC-COLING 2024*, pages 178–187, Torino, Italia. ELRA and ICCL.

David B Resnik and Mohammad Hosseini. 2024. The ethics of using artificial intelligence in scientific research: new guidance needed for a new tool. *AI and Ethics*, pages 1–23.

Quan Sun, Yufeng Cui, Xiaosong Zhang, Fan Zhang, Qiying Yu, Yueze Wang, Yongming Rao, Jingjing

Liu, Tiejun Huang, and Xinlong Wang. 2024. Generative multimodal models are in-context learners. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14398–14409.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Harish Tayyar Madabushi, Edward Gow-Smith, Carolina Scarton, and Aline Villavicencio. 2021. AStitchInLanguageModels: Dataset and methods for the exploration of idiomaticity in pre-trained language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3464–3477, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tan Wang, Kevin Lin, Linjie Li, Chung-Ching Lin, Zhengyuan Yang, Hanwang Zhang, Zicheng Liu, and Lijuan Wang. 2023. Equivariant similarity for vision-language foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11998–12008.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 641–649.

Michal Yarom, Yonatan Bitton, Soravit Changpinyo, Roee Aharoni, Jonathan Herzig, Oran Lang, Eran Ofek, and Idan Szpektor. 2024. What you see is what you read? improving text-image alignment evaluation. *Advances in Neural Information Processing Systems*, 36.

Majid Yazdani, Meghdad Farahmand, and James Henderson. 2015. Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1733–1742, Lisbon, Portugal. Association for Computational Linguistics.

Ron Yosef, Yonatan Bitton, and Dafna Shahaf. 2023. IRFL: Image recognition of figurative language. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1044–1058, Singapore. Association for Computational Linguistics.

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. 2022. When and why vision-language models behave like bags-of-words, and what to do about it? *arXiv preprint arXiv:2210.01936*.

Ziheng Zeng and Suma Bhat. 2022. Getting BART to ride the idiomatic train: Learning to represent idiomatic expressions. *Transactions of the Association for Computational Linguistics*, 10:1120–1137.

Ziheng Zeng, Kellen Cheng, Srihari Nanniyur, Jianing Zhou, and Suma Bhat. 2023. IEKG: A commonsense knowledge graph for idiomatic expressions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14243–14264, Singapore. Association for Computational Linguistics.

A Prompt Generation Steps

A.1 Subtask A: Generating Five Images

1. **Define Image Categories:**
   - **Idiomatic Meaning:** Depicts the figurative interpretation of the expression.
   - **Related to Idiomatic Meaning:** Conceptually linked but distinct from the idiomatic meaning.
   - **Related to Literal Meaning:** Loosely connected to the literal sense.
   - **Literal Meaning:** Represents the explicit, literal interpretation.
   - **Distractor:** Unrelated to both idiomatic and literal meanings.
2. **Step-by-Step Process:**
   (a) Start with clear prompts for **Idiomatic Meaning** (#1) and **Literal Meaning** (#4).
   (b) Design **Distractors** (#5) that avoid overlap with either meaning while remaining in the same category (e.g., objects, actions).
   (c) Create intermediate prompts for:
      - **Related to Idiomatic Meaning** (#2): Focus on the modifier (first word) and abstract ideas.
      - **Related to Literal Meaning** (#3): Focus on the head (second word) without overlapping with #4.
   (d) Refine and validate the sequence to ensure a smooth transition across all five images.

A.2 Subtask B: Generating Image Sequences

1. **Define Two Narratives:**
   - One sequence representing the **Idiomatic Meaning** (e.g., abstract or metaphorical concept).

- One sequence representing the **Literal Meaning** (e.g., realistic depiction of the expression).

2. **Step-by-Step Process:**
   (a) Break each meaning into three sequential steps that form a coherent narrative.
   (b) Create detailed prompts for each step:
      - For **Idiomatic Sequences:** Use concrete visualizations of abstract concepts.
      - For **Literal Sequences:** Focus on realistic and specific depictions of the literal interpretation.
   (c) Maintain consistency in visual elements (e.g., setting, characters) across the sequence for clarity.
   (d) Validate the order to ensure logical progression and refine prompts as needed.

A.3 General Tips

- Use specific, detailed descriptions to reduce ambiguity.
- Test prompts iteratively to improve quality and coherence.
- Ensure all prompts align stylistically for uniformity in the dataset.

B Example Data for Noun Compound "Bad Apple" (Subtask A)

B.1 1. Compound Information

- **Compound**: *Bad Apple*
- **Subset**: Sample
- **Sentence Type**: Idiomatic
- **Example Sentence**: *The problem for them, of course, is how to explain how these few bad apples managed to stay in place for so many years.*

B.2 2. Associated Images

Image 1

- **File Name**: `39242366111.png`
- **Caption**: The image depicts three children standing in front of a gray, textured wall. Each child is dressed in a white shirt and dark shorts, with one child wearing a backpack. The children appear to be engaged in painting or drawing on the wall.
  - **Left Child**: This child has dark hair and is holding a red spray paint can in their right hand. They are facing away from the camera, focusing on the wall. The child's left hand is holding a yellow object, possibly another piece of art equipment.
  - **Middle Child**: This child has brown hair and is also facing away from the camera. They are holding a brown spray paint can in their right hand and appear to be actively spraying it onto the wall. Their left hand is not visible, suggesting they might be holding something else out of frame.
  - **Right Child**: This child has light brown hair tied back and is also facing away from the camera. They are holding a green spray paint can in their right hand and seem to be in the process of spraying it onto the wall. Their left hand is not visible, similar to the middle child.

  The wall behind them has some graffiti already present, including the word "COOL" written in orange spray paint. There are also some additional orange squiggles and dots scattered around the graffiti, indicating that the children are in the process of adding more artwork to the wall. The overall scene suggests a playful and creative atmosphere, with the children engaged in an artistic activity.

Image 2

- **File Name**: `43074669652.png`
- **Caption**: The image depicts a cartoon-style illustration of a young boy standing at a table. The boy has spiky brown hair and is wearing an orange plaid shirt with black pants and gray shoes. His expression appears to be one of surprise or shock, as he looks towards the table in front of him. On the table, there are two cups: one is upright and filled with a dark liquid, likely coffee or tea, while the other cup is tilted and spilling its contents. The spilled liquid is causing a splash effect, with droplets and steam rising from the cup. Additionally, there are several scattered coffee beans around the table, indicating that the coffee was possibly freshly brewed and spilled.

The overall scene suggests a moment of accidental spillage, capturing the boy's reaction to the unexpected mess.

Image 3

- **File Name**: `78200848882.png`
- **Caption**: The image depicts a halved peach with a detailed and realistic appearance. The peach is split open, revealing its juicy interior. The outer skin of the peach is a vibrant orange color, with visible cracks and small white specks that resemble sugar crystals or frost. The flesh inside is a deep pinkish-orange, with a few small seeds scattered throughout. The pit, or seed, is prominently visible at the center, surrounded by a dark brown husk. A single green leaf is attached to the stem end of the peach, adding a touch of freshness to the image. The background is plain and light-colored, which helps to highlight the details and colors of the peach.

Image 4

- **File Name**: `82053252112.png`
- **Caption**: The image depicts an orange-colored apple that appears to be decomposing or decaying. The apple has a green leaf protruding from its top, indicating it was once fresh and healthy. The apple's skin is cracked and broken into several pieces, revealing the inner flesh which is also disintegrating. Small fragments of the apple's outer skin and inner flesh are scattered around the base of the apple, suggesting that the process of decay is ongoing. The apple's texture is detailed, with visible cracks and spots on its surface, adding to the realistic portrayal of decay. The background is plain white, which helps to highlight the apple and its state of decomposition.

Image 5

- **File Name**: `87462057419.png`
- **Caption**: The image depicts a rustic, burlap sack filled with several bright orange apples. The sack is tied with a simple rope and has a rough, textured appearance typical of burlap fabric. The apples are large and glossy, with a few small white specks on their surfaces, giving them a slightly speckled look. Each apple is adorned with green leaves, which add a touch of freshness to the scene. In addition to the apples inside the sack, there are three more apples placed outside the sack. Two of these apples are positioned near the bottom left and right corners of the image, while the third one is located at the bottom center. These apples also have green leaves attached to them, maintaining the consistent theme of natural elements. The background of the image is plain white, which helps to highlight the vibrant colors of the apples and the rustic texture of the burlap sack. The overall composition suggests a harvest or autumnal theme, emphasizing the abundance and freshness of the fruit.

C Example Data for Noun Compound "Bad Apple"

C.1 1. Compound Information

- **Compound**: *Bad Apple*
- **Subset**: Sample
- **Sentence Type**: Idiomatic
- **Expected Item**: `41016103905.png`

C.2 2. Sequence Captions

Sequence Caption 1 The image shows a classroom scene with five animated characters, likely children, sitting at desks. They appear to be engaged in a classroom activity, possibly a lesson or a group discussion. The classroom has a whiteboard with various colored sticky notes on it, suggesting that the students are using it for brainstorming or organizing their thoughts. There's a clock on the wall, a book on one of the desks, and a small stuffed animal on the floor. The overall atmosphere is one of a typical classroom setting.

Sequence Caption 2 The image shows an animated character, a young boy with red hair, sitting at a desk with a laptop. He is holding a kite with a star design in his hand, and the kite appears to be flying away from him. The background includes a cloud and a sun, suggesting an outdoor setting. The boy is smiling and seems to be enjoying the moment.

C.3 3. Associated Images and Captions

Image 1

- **File Name**: `09109572696.png`

- **Caption**: The image shows a group of animated characters, likely children, standing in front of a television screen that displays various cartoon animals. The characters are smiling and appear to be enjoying the show. The animals on the screen include a dinosaur, a crocodile, a fish, and a bird, suggesting that the show might be educational or entertaining, possibly teaching children about different animals. The characters are dressed in casual clothing, and the setting seems to be a living room or a similar indoor space.

Image 2

- **File Name**: `19634489503.png`
- **Caption**: The image shows an animated character, a young man with brown hair, standing next to a trash can. The trash can is overflowing with what appears to be broken pieces of a red apple, suggesting that the apple has been smashed or shattered. The character is holding his hand out as if he is about to catch or pick up one of the apple pieces. The scene is stylized and cartoonish, with exaggerated features and a limited color palette.

Image 3

- **File Name**: `39830945702.png`
- **Caption**: The image shows a cartoon illustration of a young girl with red hair, wearing a white sweater with a red bow and a plaid skirt. She is sitting at a small desk and appears to be reading a book. In front of her is a wooden desk with a chair, and on the desk is a red apple with a green leaf. The girl is smiling and seems to be enjoying her time reading. The background is plain white, which puts the focus on the girl and her activity.

Image 4

- **File Name**: `41016103905.png`
- **Caption**: The image shows a group of animated characters that appear to be in a state of distress or chaos. They are depicted with exaggerated expressions and body language, suggesting a scene of panic or fear. The characters are styled in a cartoonish manner, with a limited color palette that gives the image a somewhat muted and gritty look. The background has splatter effects that add to the sense of disarray. The characters are not clearly identifiable, but they seem to be in a room with a desk and a chair, which might suggest a home or office setting.

C.4 4. Phrase Context and Meaning

- **Phrase Context**: *One bad apple can spoil the atmosphere of an otherwise positive workplace.*
- **Meaning**: A person who has a negative influence on others or spoils a group due to their behavior or character.