PuzzleGPT: Emulating Human Puzzle-Solving Ability for Time and Location Prediction

Anonymous ACL submission

Abstract

The task of predicting time and location from images is challenging and requires complex human-like puzzle-solving ability over different clues. In this work, we formalize this ability into core skills and implement them using different modules in an expert pipeline called PuzzleGPT. PuzzleGPT consists of a perceiver to identify visual clues, a reasoner to deduce prediction candidates, a combiner to combinatorially combine information from different clues, a web retriever to get external knowledge if the task can't be solved locally, and a noise filter for robustness. This results in a zero-shot, interpretable, and robust approach that records state-of-the-art performance on two datasets – TARA and WikiTilo. PuzzleGPT outperforms large VLMs such as BLIP-2, InstructBLIP, LLaVA, and even GPT-4o, as well as automatically generated reasoning pipelines like VisProg(Gupta and Kembhavi, 2022), by at least 32% and 38%, respectively. It even rivals or surpasses fine-tuned models.

1 Introduction

Recent advances in Vision-Language (VL) research have led to models that perform impressively (Zhu et al., 2023; Li et al., 2022; Lu et al., 2022a; Alayrac et al., 2022; Team et al., 2024) on a variety of tasks such as GQA (Hudson and Manning, 2019), VQA v2 (Antol et al., 2015), VCR (Zellers et al., 2019), OK-VQA (Marino et al., 2019), ScienceQA (Lu et al., 2022b), visual entailment (Xie et al., 2019). Chain-of-thought reasoning (Lu et al., 2024, 2022b). These tasks primarily assess one of, or at most a combination of, perception, reasoning, and outside knowledge retrieval abilities. For example, OK-VQA requires perception and outside knowledge retrieval, and GQA and VCR require perception and commonsense reasoning.

However, humans seamlessly integrate a variety of skills – perception, reasoning, knowledge retrieval, and common sense – to solve complex, multi-step problems. Tasks and benchmarks that test these diverse skills are crucial for developing models that approach human-level reasoning. The task of time and place reasoning from images proposed by TARA (Fu et al., 2022) takes a step closer to this goal. It demands a mix of perception, reasoning, combinatorial, and outside knowledge retrieval abilities over multiple steps. It is like solving a puzzle. For example, in Figure 1, it is required to detect entities such as Times Square and visual text, "Justice for George Floyd". Then, a reasoner needs to deduce possible location (New York/United States) and time candidates (post-2000, 2020-2021) from these clues. Next, these candidates need to be combined in various ways to find a candidate at the intersection of all candidates (post 2000 $\cap$ 2020-2021 = 2020-2021). Finally, if the answer is still unclear (2020-2021), a web search is required using the deduced information.

The practical applications of the task stem from its focus on images depicting events that occurred at a specific location and time. This has incredibly useful applications, such as timeline construction, and stitching together news stories from online pictures and social media posts.

Existing works take a direct approach to this nuanced problem. TARA tries to supervise a model directly to predict location and time directly. The hope is that the model will learn to identify time and location discriminative clues implicitly, given appropriate supervisory signals. While the approach might have worked for a limited time and location candidates, the real scenario of hundreds of locations/time with fine differences makes this approach unscalable, and thus impractical. QR-CLIP (Shi et al., 2023) additionally tries to incorporate external knowledge in the learning process. However, it oversimplifies the problem and assumes mere retrieval can accomplish the task without relying on specific clues and their combinatorial intersections.

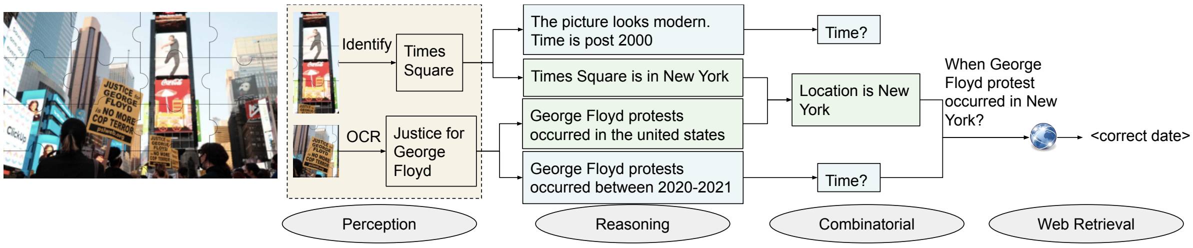

Figure 1: Figure describing the complexity of the TARA task and our approach to it.

We argue that a complex puzzle-like problem like TARA, requires an equally well thought-out solution. To this end, we propose PuzzleGPT. PuzzleGPT abstracts the skills required to solve it into five core abstract ideas: perceiver, reasoner, combiner, noise filter, and knowledge retriever. It represents each with specific modules that perform specific tasks. The perceiver processes visual signals and identifies different entities such as people, buildings, cultural signals, and OCR text. For each of these entities, the reasoner tries to deduce their co-relations with a location and time candidate. Integrating clues from multiple entities is crucial for accurate prediction. However, simply combining all clues can introduce noise from irrelevant information, while relying on individual clues might provide insufficient context. To address this challenge, we propose a confidence-based hierarchical combination approach. This approach analyzes clues at increasing levels of granularity: first, individual entities; then, pairs; followed by triplets, and so on, tracking candidate predictions. The process stops once a candidate reaches a threshold vote, efficiently combining entities while minimizing noise.

Apart from being zero-shot, our design choices lend PuzzleGPT desirable properties. Reasoner makes the approach interpretable and thus trustworthy. A hierarchical combination approach makes it not only combinatorial but also noise-resistant. Web retriever infuses the approach with the ability to incorporate world knowledge into the reasoning process. The noise filter adds further robustness.

PuzzleGPT scores state-of-the-art (SOTA) zero-shot performance on TARA, coming close to or surpassing even fine-tuned approaches. We demonstrate that our method outperforms existing SOTA VL models like Instruct BLIP (Dai et al., 2023), BLIP-2 (Li et al., 2023), LLaVA (Liu et al., 2023a), by a margin of at least 32% (standardized location accuracy). It even beats the popular proprietary GPT-4o(OpenAI, 2024). This highlights current VLMs' inability to simultaneously employ multiple skills to solve a task. We also report superior performance to automatically generated modular pipelines like VisProg, indicating generating an automatic pipeline for this complex task perhaps exceeds their current capabilities. Furthermore, we show that our method generalizes and scores SOTA on another location and time dataset, WikiTilo (Zhang et al., 2024).

We make the following contributions:

- We propose a novel method, PuzzleGPT, to emulate human puzzle-solving ability for predicting time and location from images.
- Our design choices make our approach interpretable, robust, combinatorial, and retrieval augmented.
- PuzzleGPT scores SOTA performance on TARA and WikiTilo.

2 Related Work

Vision-Language Models. Recently, VLMs (Radford et al., 2021; Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023b) have demonstrated remarkable multimodal capabilities through large-scale vision-language training. One family of VLMs such as CLIP (Radford et al., 2021) typically trains a visual encoder and a text encoder to map visual and text input into a common embedding space. The resulting visual encoders are widely adopted to extract visual features that are fed to LLMs in the other family of work (Alayrac et al., 2022; Li et al., 2023; Liu et al., 2023b; Lin et al., 2023). For example, LLaVA takes CLIP's visual feature as input and is trained to generate target text. These VLMs with text-generation abilities have achieved superior performance on vision-language datasets.

Visual Reasoning Datasets. Early work like VQA (Antol et al., 2015) mainly probes perception more than reasoning abilities, while datasets like CLEVR (Johnson et al., 2017) focused on compositional reasoning in a controlled synthetic en-

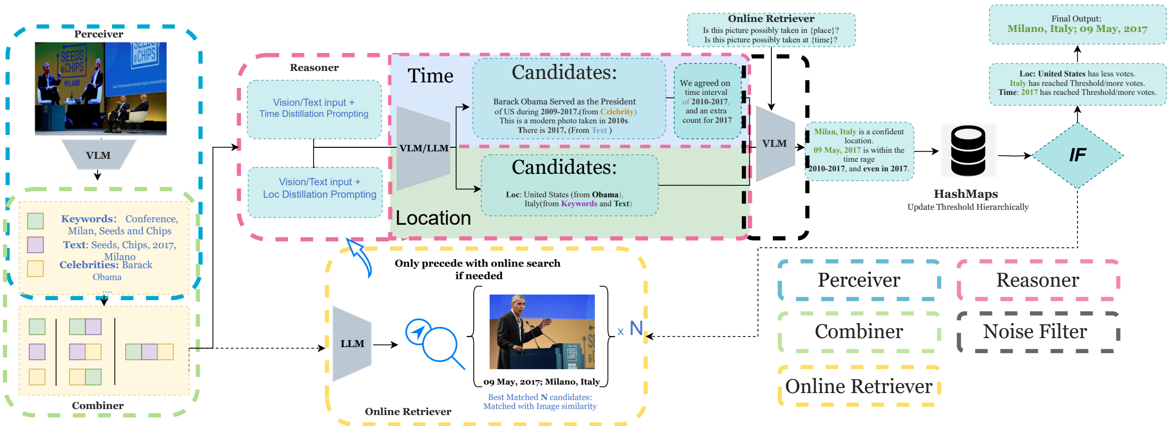

Figure 2: Model Overview. All VLMs/LLMs are pretrained and frozen. The online retriever will be accessed only if the model is not confident about existing results.

vironment. GQA (Hudson and Manning, 2019) pushed towards scene understanding with structured knowledge graphs. Recent work further tackles visual reasoning from different aspects (Zellers et al., 2019; Han et al., 2023; Fu et al., 2024). However, these datasets either do not require multiple steps of reasoning or lack the depth and breadth of required knowledge. TARA (Fu et al., 2022) and WikiTiLo (Zhang et al., 2024), on the other hand, necessitates multi-step, puzzle-like reasoning over multiple visual clues, combined with external knowledge, posing a unique challenge for existing VLMs. The performance of VLMs such as LLaVA and BLIP-2 remains unsatisfactory on these two datasets. A recent retrieval-based supervised method (Shi et al., 2023) is proposed to augment CLIP with world knowledge, but it does not yield significant advancement on these tasks either. More importantly, these retrieval-based models' predictions are difficult to interpret.

Neural Program Induction / Modular Networks. Inspired by the need for more composable and interpretable models, research in neural program induction aims to learn programs or modules for solving tasks. Early work explored differentiable neural programmer (Neelakantan et al., 2015), while Neural Module Networks (Andreas et al., 2016) focused on composing visual modules for reasoning. More recently, VisProg(Gupta and Kembhavi, 2022) proposed automatic code generation for VQA tasks. However, as our experiments show, automatically generating effective pipelines for intricate problems like TARA remains challenging. PuzzleGPT's expert-designed pipeline, tailored specifically for time and location puzzle-solving, outperforms these automatic approaches, suggesting the importance of domain knowledge and task-specific design for complex reasoning.

3 Methodology

We propose PuzzleGPT to emulate human-like puzzle-solving ability. It consists of an expert pipeline consisting of specific modules that represent distinct skills, as illustrated in Figure 2. In this section, we describe each of the modules in detail.

Perceiver. Perceiver (denoted as $\mathcal{P}$) processes visual signals. Given an image, PuzzleGPT will first scan the image to find entities of interest, such as celebrities, text, landmarks, or other types of keywords. By finding the entities, the Perceiver can focus on patches containing specific entities and reasoning independently. This enables it to generate specific textual knowledge about the entities (for a landmark, its name; for text, its Optical Character Recognition; and so on). We use BLIP-2 as the Perceiver in this work.

Reasoner. Reasoner in PuzzleGPT is an LLM that deduces time/location clues from Text Knowledge produced by Perceiver. An example is shown in Figure 2. Based on the presidential term, the reasoner can target a time range for TEXT <Barack Obama> as 2009-2017. In addition, based on words that appear on the image, it can also recognize "Milano" as a location clue. Given the GPT models' impressive reasoning abilities, we use GPT-3.5 as a Reasoner in this work.

Combiner. While perceiving and reasoning entities independently might provide a larger search space, there is a need to detect the connections across different entities. For instance, in Figure 2, reasoning based on the celebrity name may suggest the location candidate as the United States, even though text clues suggested “Milano". Therefore, we construct a combining strategy to divide available information into three hierarchies: the first hierarchy will reason independently, the second hierarchy will reason based on a combination of information from a pair of independent entity sources, and the third hierarchy will work based on a combination of all available entities. Three hierarchies is not required to be fixed, but is a design choice for efficiency and computational limitations. This strategy significantly enlarged the possibility of finding a targeting combination of knowledge that maximizes the recall of extracting time/location clues.

Noise Filter. In the hierarchical combiner, we enlarged the search space size to find appropriate clue combinations. However, hierarchical combinations will also bring erroneous combinations. Erroneous information will not benefit the reasoning process and can even introduce a significant bias. To address such bias introduction, we employed a VLM to decide whether the candidate voted by the reasoner is a “Real Candidate", based on its background knowledge. We use BLIP-2 as a Noise Filter as well.

Online Retriever. VLM/LLMs are, at times, insufficient to reason complicated tasks based only on static knowledge priors obtained through pertaining. From another perspective, human will access online resources once their knowledge is insufficient. To mimic such an information augmentation for puzzling solving, we allow the model to generate a search query through the Reasoner by providing evidence combination from the combiner. Then, it accesses online dynamic resources through a web search engine. To reduce noise, the online retriever evaluates the relevance between retrievals and the original image through image-to-image/text similarities. Only retrievals scoring higher than a Retrieval Threshold (RT), are kept. The retrievals are then fed to the Reasoner to extract the candidate time/location. We use CLIP to generate retrieval scores.

3.1 Risk Mitigation

As PuzzleGPT’s design can generate and obtain significant knowledge and information, it is exposed to a lot of noise. They can originate from poor perception, hallucination or poor web retrievals. This

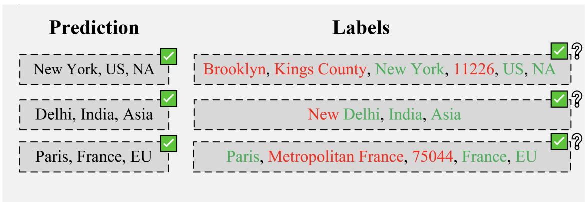

Figure 3: Unstructured location labels lead to unfair comparison for exact match Accuracy metric. We mitigate this by label standardization.

needs mitigation. While Noise Filter aids towards this step, it’s not sufficient.

To this end, instead of finding one specific location/time candidate, we instead try to find the location/time candidate **with highest confidence** hierarchically. That is, we maintained two hash maps for location and time reasoning, each of which records a candidate accepted by the noise filter hierarchically. By hierarchical, the hashmap will update different hierarchies of a candidate separately. For instance, if PuzzleGPT collects <New York, US, NA> and it is accepted by the noise filter, then <NA> will be first recorded in the **continent** hashmap, along with <US> being updated in the **country** hashmap under **continent** <NA>, and the same for **city** hierarchy <New York>. A similar strategy is applied to Time updating too, with different hierarchies being Year, Month, and Day. To define the state of being ’confident’, We set a hyperparameter Hash Threshold, denoted as HT and initially set to 5. If HT is reached by a candidate, then we know the system would be confident enough that this candidate is the correct answer, and the reasoning process, whichever stage it stands, will (early) stop. If the threshold is never reached, the candidate with the highest confidence will be the output, representing our most confident answer. The hash threshold HT is initially set to 5.

4 Experiments

In this section, we report our results on two datasets: TARA and WikiTilo.

4.1 TARA

TARA is sourced from the New York Times and requires time and location prediction for images depicting newsworthy events. In total, there are around 1.5K samples in the test and validation set. The label set is open-ended with a unique label for each sample.

4.1.1 Metric

The open-ended nature of labels in TARA makes evaluation challenging. Two metrics were proposed originally – Accuracy and Example-F1.

Accuracy measures the exact match of the prediction with the label. While this works for time evaluation where the labels are properly formatted (YYYY-MM-DD), location evaluation leads to unreliable results as the labels are highly unstructured. As illustrated in Fig Figure 3. in addition to city, country, and country, some labels contain additional information such as Pin Code, county name (Kings County), and geographical area name (Metropolitan France). This causes even correct predictions to be incorrectly classified as wrong. To address this, we standardize all locations into city, country, and continent using GeoPy [1] into city, country, and country. Further, if the label contains a specific area within a city (e.g. Times Square or Central Park), we keep that to not lose location precision. We use these formatted labels for measuring accuracy and call it Standardized Accuracy (Std. Acc).

To measure partial correctness – only correct year or only correct continent and country – TARA proposes Example-F1 metric. It is defined as follows:

$$\text{Example-F1} \; = \frac{2|GT \cap Pred|}{|GT| + |Pred|}$$

As the score is inversely dependent on $|Pred|$, shorter predictions are unduly rewarded. For example, a model that predicts only year scores abnormally high Example-F1. We mitigate this bias by adding a brevity penalty, following NLP literature (Papineni et al., 2002):

$$\text{Example-F1}^{\beta} = e^{-(\frac{|Pred|}{|GT|}-1)^{+}} \cdot \text{Example-F1}$$

We use Example-F1 and F1 interchangeably in this work from here onwards.

4.1.2 Baselines

In addition to comparing PuzzleGPT against previously reported approaches on TARA, we also evaluate it against recent VLMs to provide a comprehensive comparison and valuable insights.

Large Vision Language Models. We evaluate the following VLMs: BLIP-2, InstructBLIP, LLaVA, and GPT-4o. These models leverage the power of LLMs for visual reasoning, thereby acquiring extensive knowledge and reasoning abilities. They represent single-stop solutions for complex tasks.

[1] https://geopy.readthedocs.io/en/stable/

Model	Time Acc(%)	Time F1$^{\beta}$	Location Std. Acc(%)	Location F1$^{\beta}$
BLIP2	0.30	32.27	*17.41*	43.59
LLaVA	0.23	*43.26*	7.85	25.92
GPT4o	0.30	21.94	16.62	*47.16*
InstructBLIP	0.00	33.83	16.69	26.05
IdealGPT	0.27	26.83	9.95	25.70
VisProg	0.00	18.52	0.00	4.74
PuzzleGPT	**0.30**	**43.72**	**22.99**	**51.04**

Table 1: We compare PuzzleGPT to SOTA zero-shot generative VLMs on TARA. PuzzleGPT outperforms all prior methods, scoring SOTA performance.

Model	Time Acc(%)	Time F1	Location Acc(%)	Location F1
CLIP	0.46	39.90	11.11	44.96
CLIP+	1.00	43.09	15.72	49.74
CLIP+Seg	0.92	42.82	16.46	50.52
QR-CLIP	**3.53**	**47.89**	19.31	50.96
PuzzleGPT	0.30	*43.72*	**22.99***	**56.11**

Table 2: PuzzleGPT comparison against representative classification models reported in prior works. All are finetuned except CLIP. * denotes Std. Acc. PuzzleGPT outperforms finetuned methods on location reasoning while recording comparable performance on time prediction.

Code Based Modular Approaches. We also compare PuzzleGPT to methods that generate modular code for various VL tasks, such as VisProg. These methods serve as references for automatic pipelines, contrasting with our expert pipeline. Additionally, we compare against IdealGPT, which aims to enhance robustness in automatic pipelines through an iterative pipeline.

4.1.3 Results

We compare PuzzleGPT against zero-shot VLMs in Table 1 and finetuned approaches in Table 2. We make the following observations:

PuzzleGPT records state-of-the-art performance. PuzzleGPT outperforms all methods, including the popular GPT-4o model, for both location and time prediction. It's especially skilled at location prediction: >30% Std. Acc. improvement over next best method (BLIP-2).

PuzzleGPT is more skilled than single-stop Large VLMs. PuzzleGPT's strong improvements over all VLMs indicate their limitation in leverag-

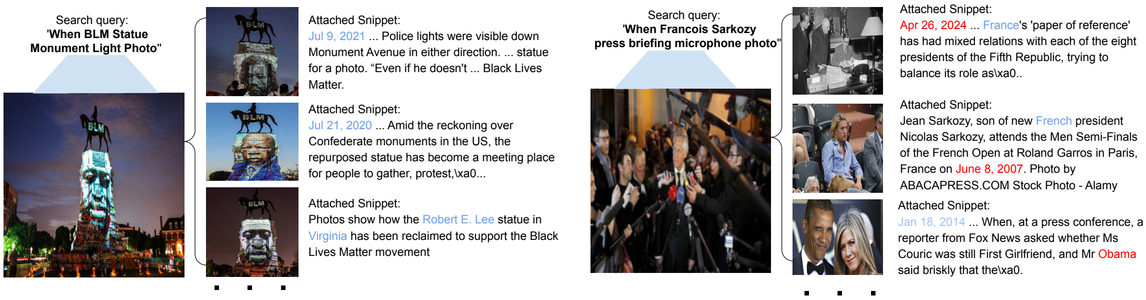

Figure 4: With specific and clear clues, our model can retrieve high-quality web content while generic images tend to retrieve noisy content.

Ablations	Time-F1$^{\beta}$	Location-F1$^{\beta}$
PuzzleGPT	**43.72**	**51.04**
- w/o Filtering	39.27	48.77
- w/o Retrieval	42.63	43.30

Table 3: Noise Filter and Retriever ablation. Performance drop if we remove either of them, underscoring their importance to PuzzleGPT.

Ablations	Time-F1$^{\beta}$	Location-F1$^{\beta}$
PuzzleGPT (I-I Retrieval)	**43.72**	**51.04**
- I-T Retrieval	43.47	50.95

Table 4: I: Image. T: Text. Retrieval is best served by image-image matching. Replacing it with image-text retrieval reduces performance.

Ablations	Time-F1$^{\beta}$	Location-F1$^{\beta}$
PuzzleGPT	**43.72**	**51.04**
1st Hier Only	42.62	46.68
3rd Hier Only	42.72	45.21

Table 5: Hier: Hierarchy. Confidence-based hierarchical combination is critical. PuzzleGPT outperforms simpler methods by avoiding incomplete information from 1st Hier Only and noise from 3rd Hier Only.

ing diverse skills to accomplish this complex task.

PuzzleGPT's expert pipeline is better at puzzle-like tasks than automatic pipelines. From Visprog's inferior performance, we conclude that automatic pipelines are 1) constrained by the types of skills they can apply and 2) the search space for the optimum pipeline in puzzle-like tasks is so large that they generate suboptimal code.

PuzzleGPT comes close to or surpasses fine-tuned performance. PuzzleGPT's effectiveness and strong performance are highlighted by the fact that it achieves >10% Example-F1 improvement over the best fine-tuned approach (QR-CLIP).

4.1.4 Ablation Studies

We investigate PuzzleGPT from different axes to thoroughly analyze its modules.

Confidence-based hierarchical combination is crucial. To understand the importance of hierarchical combination, we compare our approach in Table 5 to simple ablations that 1) do not combine information from entities (1st Hier Only), and 2) combine information from all entities in one go (3rd Hier Only). PuzzleGPT outperforms both. Figure 5 illustrates the underlying reason: 1st Hier only results in incomplete information and 3rd Hier is noisy. These results demonstrate that a confidence-based hierarchical combination is crucial to carve a middle path between incorporating signals from different puzzle pieces and reducing noise.

Confidence thresholding matters in hierarchical combination. Figure 6 shows that the best performance is reached at threshold=90, with inferior scores for both lower and higher thresholds. This implies low confidence threshold allows noisy candidates to be predicted, while a higher threshold results in more pipeline iterations, thereby introducing additional noisy candidates.

Web retrieval augments PuzzleGPT with external knowledge. As reported in Table 3, not retrieving external knowledge from the internet leads to a performance drop of 1.09% in Time-F1$^{\beta}$ and 7.74% in Location-F1$^{\beta}$. Figure 4 further illustrates the importance of retrieval, especially for time prediction.

Retrieval is sensitive to thresholding. Figure 6 plots the model performance against different values of retrieval threshold. The performance peaks at 90, implying the lower threshold is noisy and the higher threshold leads to information bottleneck.

Retrieval is best served by image-image matching. Table 4 reports the performance achieved by replacing image-image retrieval with image-text retrieval. We find that it leads to a performance drop of 0.25% in Time-F1$^{\beta}$, and 0.09% in Location$^{\beta}$, indicating that it's a suboptimal strategy for this task.

Time prediction is more complex than location

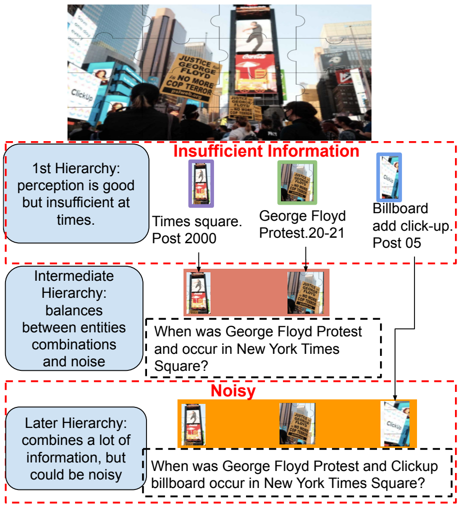

Figure 5: An illustration of a hierarchical combination of information. Images with different colored images mean they are processed separately in the hierarchy, for example, all three images are processed together in the final hierarchy. 1st Hierarchy leads to a scarcity of information, while 3rd Hierarchy is noisy. This underscores the need for a hierarchical combination.

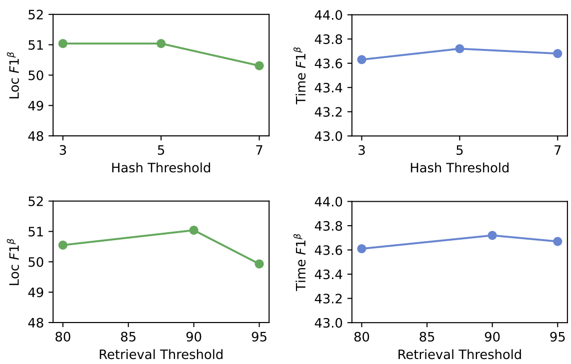

Figure 6: Top: Ablation on Hash Threshold (HT): peak performance at 5, with noisy performance on both lower or higher HT. Bottom: Ablation on Retrieval Threshold (RT): retrieval is best at 90, with either side of it leading to noisy retrieval.

prediction. We observe from Figure 7 that the majority of queries for location finish in the first hierarchy, while almost all queries for time reach the third hierarchy. This demonstrates that location prediction is doable from individual visual clues, while time prediction requires more combinations of clues to arrive at a candidate. Further, Figure 8 reveals that almost all queries for time need web retrieval. All this points to a higher complexity of time prediction.

Noise filtering is critical. From Table 3, we observe that eliminating the noise filtering module from PuzzleGPT leads to a performance drop of 4.45% in Time-F1$^{\beta}$ and 2.27% in Location-F1$^{\beta}$. This highlights the importance of noise filtering.

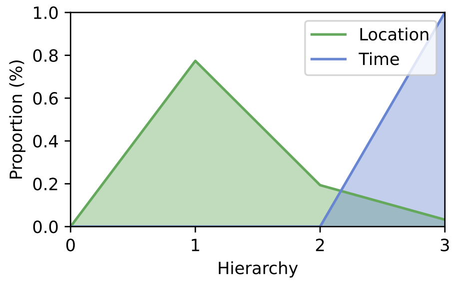

Figure 7: Distribution of endpoints of hierarchy. Time prediction is more complex than location prediction, with most location queries resolved in the first hierarchy while most time queries reach the third hierarchy.

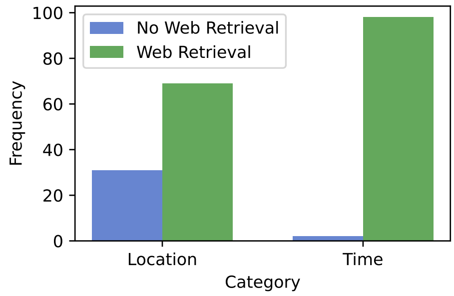

Figure 8: Frequency of retrieval and no retrieval between time and location queries. Almost all time queries require web retrieval, highlighting the complexity of time prediction compared to location prediction.

4.1.5 Qualitative Analysis

We conducted case studies on TARA for qualitative analysis to further demonstrate PuzzleGPT's effectiveness. Recording the instances and the reasoning steps, we select a positive cases to showcase that our model was able to capture information and deliver reliable inference even from noisy resources such as the internet (See Figure 9). Meanwhile, we also noticed that in some cases, especially for images that contain less clear/informative clues, our model can fail to discover and ground time clues. We report them in the Appendix. In general, if only general contexts, such as image caption and event description, are available, the search query generated can also be too general to properly search the web. We also noticed that, even for challenging images (available in the Appendix), PuzzleGPT is still able to increment confidence about the correct location candidate. This is consistent with the situation that PuzzleGPT performed better on location scores.

4.2 WikiTilo

To demonstrate the generalization of our approach, we also report PuzzleGPT's performance on another location/time reasoning dataset, WikiTilo. It contains ~600 images in the test set with a fo-

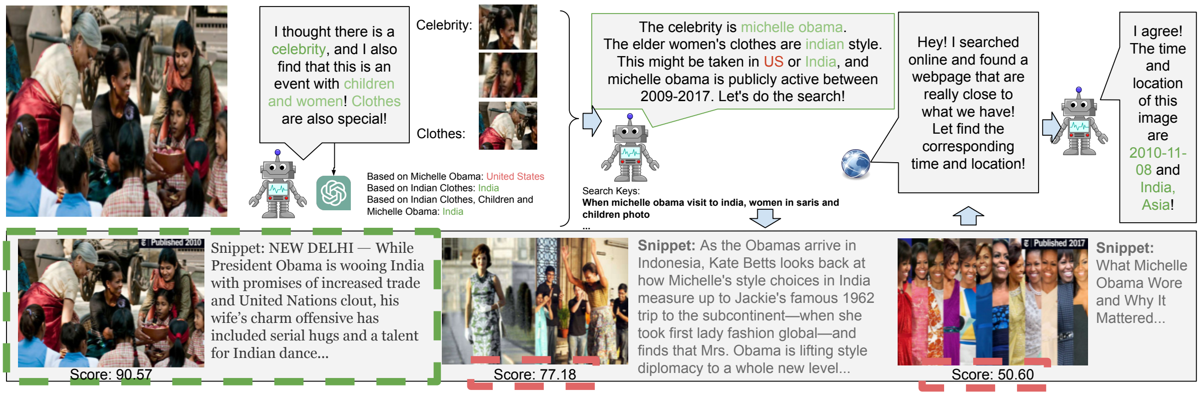

Figure 9: We showcase one positive case while implementing PuzzleGPT. More positive and negative samples are available in the supplementary section.

	Time			Country			Region		
Models	**Acc(%)**	**Prec**	**F1**	**Acc(%)**	**Prec**	**F1**	**Acc(%)**	**Prec**	**F1**
OpenFlamingo Test	27.70	26.36	11.49	3.89	3.69	2.18	4.72	8.62	4.72
OpenFlamingo-VQA	31.59	30.36	28.60	**48.88**	*53.78*	*41.19*	22.49	30.49	18.64
OpenFlamingo-VQA CoT	35.21	29.36	28.42	40.3	45.24	33.17	*24.04*	*39.39*	*19.27*
LLaMA-AdapterV2-Instr[a]	*58.02*	28.04	32.88	23.05	52.64	18.66	19.07	26.59	13.01
LLaMA-AdapterV2-Instr[b]	34.34	*58.59*	*37.70*	*45.62*	51.57	35.50	11.12	10.05	5.99
Frequency baseline	25.07	25.29	23.27	3.33	2.95	2.88	12.53	12.59	12.25
PuzzleGPT(Ours)	**71.90**	**70.63**	**72.61**	43.65	**72.78**	**49.79**	**62.06**	**79.22**	**68.18**

Table 6: PuzzleGPT generalizes to WikiTilo dataset, scoring state-of-the-art performance in almost all the metrics.

cus on identifying sociocultural cues to predict time/location. Whereas TARA evaluates predictions on open-ended labels, WikiTilo's labels are multi-choice. For location, the evaluation is divided into Country, with 30 multiple choice labels, and Region, with 8 distinct labels. For a time, the labels are divided into four time periods Since the labels are multi-choice, the predictions are simply scored on accuracy, precision, and F_1 score.

As reported in Table 6, we score state-of-the-art performance on WikiTilo for time and region prediction. Specifically, our method improves time Acc. and F1. by $+23.9\%$ and $+123.5\%$ respectively, over the next best method. Region Acc. and F1. are boosted by $+158.2\%$ and $+101.1\%$ respectively. For country prediction, our Acc is slightly worse (-10.7%), but we still outperform the previous best F1 by $+68.3\%$. In contrast to our approach, previous methods fail to align countries with regions and display inconsistent behavior by scoring higher on Region (8 categories) than on Country (30 categories). We conclude from this that our approach could be a solid alternative for reducing inconsistencies and hallucinations.

Indeed, PuzzleGPT predicts time much more accurately on WikiTilo than on TARA, indicating that time prediction on TARA could be unusually challenging.

5 Conclusion

This work proposes an iterative puzzle-solving method - PuzzleGPT that consistently outperforms current SOTA VLMs on TARA, as shown by extensive experiments. We believe PuzzleGPT can further push the boundary of the current progress of VLU and point to an underexplored direction for future development.

Limitations

While PuzzleGPT demonstrates strong performance on time and location prediction tasks like TARA and WikiTilo, it's important to acknowledge its limitations. The model's architecture is specifically tailored for puzzle-like reasoning scenarios, and its performance on tasks with different structures or knowledge requirements remains unexplored. Furthermore, the current reliance on GPT for reasoning introduces dependencies on proprietary models, potentially limiting accessibility and introducing inherent biases. Future work will explore alternative reasoning modules and evaluate PuzzleGPT's generalization ability across diverse visual reasoning tasks.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 39–48.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: Visual Question Answering. In *International Conference on Computer Vision (ICCV)*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Preprint*, arXiv:2305.06500.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. *arXiv preprint arXiv:2404.12390*.

Xingyu Fu, Ben Zhou, Ishaan Preetam Chandratreya, Carl Vondrick, and Dan Roth. 2022. There is a time and place for reasoning beyond the image. *Preprint*, arXiv:2203.00758.

Tanmay Gupta and Aniruddha Kembhavi. 2022. Visual programming: Compositional visual reasoning without training. *Preprint*, arXiv:2211.11559.

Xiaotian Han, Quanzeng You, Yongfei Liu, Wentao Chen, Huangjie Zheng, Khalil Mrini, Xudong Lin, Yiqi Wang, Bohan Zhai, Jianbo Yuan, et al. 2023. Infimm-eval: Complex open-ended reasoning evaluation for multi-modal large language models. *arXiv e-prints*, pages arXiv–2311.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. *Preprint*, arXiv:1902.09506.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, Hehong Chen, Guohai Xu, Zheng Cao, Ji Zhang, Songfang Huang, Fei Huang, Jingren Zhou, and Luo Si. 2022. mplug: Effective and efficient vision-language learning by cross-modal skip-connections. *Preprint*, arXiv:2205.12005.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Xudong Lin, Simran Tiwari, Shiyuan Huang, Manling Li, Mike Zheng Shou, Heng Ji, and Shih-Fu Chang. 2023. Towards fast adaptation of pretrained contrastive models for multi-channel video-language retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14846–14855.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023b. Visual instruction tuning. *Preprint*, arXiv:2304.08485.

Jiasen Lu, Christopher Clark, Rowan Zellers, Roozbeh Mottaghi, and Aniruddha Kembhavi. 2022a. Unified-io: A unified model for vision, language, and multi-modal tasks. *Preprint*, arXiv:2206.08916.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *Preprint*, arXiv:2310.02255.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022b. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Conference on Computer Vision and Pattern Recognition (CVPR)*.

Arvind Neelakantan, Quoc V Le, and Ilya Sutskever. 2015. Neural programmer: Inducing latent programs with gradient descent. *arXiv preprint arXiv:1511.04834*.

OpenAI. 2024. Gpt-4o. Accessed: 2024-06-16.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. *Preprint*, arXiv:2103.00020.

Weimin Shi, Mingchen Zhuge, Dehong Gao, Zhong Zhou, Ming-Ming Cheng, and Deng-Ping Fan. 2023. Qr-clip: Introducing explicit open-world knowledge for location and time reasoning. *Preprint*, arXiv:2302.00952.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M. Dai, Anja Hauth, Katie Millican, David Silver, Melvin Johnson, Ioannis Antonoglou, Julian Schrittwieser, Amelia Glaese, Jilin Chen, Emily Pitler, Timothy Lillicrap, Angeliki Lazaridou, Orhan Firat, James Molloy, Michael Isard, Paul R. Barham, Tom Hennigan, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, Ryan Doherty, Eli Collins, Clemens Meyer, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Jack Krawczyk, Cosmo Du, Ed Chi, Heng-Tze Cheng, Eric Ni, Purvi Shah, Patrick Kane, Betty Chan, Manaal Faruqui, Aliaksei Severyn, Hanzhao Lin, YaGuang Li, Yong Cheng, Abe Ittycheriah, Mahdis Mahdieh, Mia Chen, Pei Sun, Dustin Tran, Sumit Bagri, Balaji Lakshminarayanan, Jeremiah Liu, Andras Orban, Fabian Güra, Hao Zhou, Xinying Song, Aurelien Boffy, Harish Ganapathy, Steven Zheng, HyunJeong Choe, Ágoston Weisz, Tao Zhu, Yifeng Lu, Siddharth Gopal, Jarrod Kahn, Maciej Kula, Jeff Pitman, Rushin Shah, Emanuel Taropa, Majd Al Merey, Martin Baeuml, Zhifeng Chen, Laurent El Shafey, Yujing Zhang, Olcan Sercinoglu, George Tucker, Enrique Piqueras, Maxim Krikun, Iain Barr, Nikolay Savinov, Ivo Danihelka, Becca Roelofs, Anaïs White, Anders Andreassen, Tamara von Glehn, Lakshman Yagati, Mehran Kazemi, Lucas Gonzalez, Misha Khalman, Jakub Sygnowski, Alexandre Frechette, Charlotte Smith, Laura Culp, Lev Proleev, Yi Luan, Xi Chen, James Lottes, Nathan Schucher, Federico Lebron, Alban Rrustemi, Natalie Clay, Phil Crone, Tomas Kocisky, Jeffrey Zhao, Bartek Perz, Dian Yu, Heidi Howard, Adam Bloniarz, Jack W. Rae, Han Lu, Laurent Sifre, Marcello Maggioni, Fred Alcober, Dan Garrette, Megan Barnes, Shantanu Thakoor, Jacob Austin, Gabriel Barth-Maron, William Wong, Rishabh Joshi, Rahma Chaabouni, Deeni Fatiha, Arun Ahuja, Gaurav Singh Tomar, Evan Senter, Martin Chadwick, Ilya Kornakov, Nithya Attaluri, Iñaki Iturrate, Ruibo Liu, Yunxuan Li, Sarah Cogan, Jeremy Chen, Chao Jia, Chenjie Gu, Qiao Zhang, Jordan Grimstad, Ale Jakse Hartman, Xavier Garcia, Thanumalayan Sankaranarayana Pillai, Jacob Devlin, Michael Laskin, Diego de Las Casas, Dasha Valter, Connie Tao, Lorenzo Blanco, Adrià Puigdomènech Badia, David Reitter, Mianna Chen, Jenny Brennan, Clara Rivera, Sergey Brin, Shariq Iqbal, Gabriela Surita, Jane Labanowski, Abhi Rao, Stephanie Winkler, Emilio Parisotto, Yiming Gu, Kate Olszewska, Ravi Addanki, Antoine Miech, Annie Louis, Denis Teplyashin, Geoff Brown, Elliot Catt, Jan Balaguer, Jackie Xiang, Pidong Wang, Zoe Ashwood, Anton Briukhov, Albert Webson, Sanjay Ganapathy, Smit Sanghavi, Ajay Kannan, Ming-Wei Chang, Axel Stjerngren, Josip Djolonga, Yuting Sun, Ankur Bapna, Matthew Aitchison, Pedram Pejman, Henryk Michalewski, Tianhe Yu, Cindy Wang, Juliette Love, Junwhan Ahn, Dawn Bloxwich, Kehang Han, Peter Humphreys, Thibault Sellam, James Bradbury, Varun Godbole, Sina Samangooei, Bogdan Damoc, Alex Kaskasoli, Sébastien M. R. Arnold, Vijay Vasudevan, Shubham Agrawal, Jason Riesa, Dmitry Lepikhin, Richard Tanburn, Srivatsan Srinivasan, Hyeontaek Lim, Sarah Hodkinson, Pranav Shyam, Johan Ferret, Steven Hand, Ankush Garg, Tom Le Paine, Jian Li, Yujia Li, Minh Giang, Alexander Neitz, Zaheer Abbas, Sarah York, Machel Reid, Elizabeth Cole, Aakanksha Chowdhery, Dipanjan Das, Dominika Rogozińska, Vitaliy Nikolaev, Pablo Sprechmann, Zachary Nado, Lukas Zilka, Flavien Prost, Luheng He, Marianne Monteiro, Gaurav Mishra, Chris Welty, Josh Newlan, Dawei Jia, Miltiadis Allamanis, Clara Huiyi Hu, Raoul de Liedekerke, Justin Gilmer, Carl Saroufim, Shruti Rijhwani, Shaobo Hou, Disha Shrivastava, Anirudh Baddepudi, Alex Goldin, Adnan Ozturel, Albin Cassirer, Yunhan Xu, Daniel Sohn, Devendra Sachan, Reinald Kim Amplayo, Craig Swanson, Dessie Petrova, Shashi Narayan, Arthur Guez, Siddhartha Brahma, Jessica Landon, Miteyan Patel, Ruizhe Zhao, Kevin Villela, Luyu Wang, Wenhao Jia, Matthew Rahtz, Mai Giménez, Legg Yeung, James Keeling, Petko Georgiev, Diana Mincu, Boxi Wu, Salem Haykal, Rachel Saputro, Kiran Vodrahalli, James Qin, Zeynep Cankara, Abhanshu Sharma, Nick Fernando, Will Hawkins, Behnam Neyshabur, Solomon Kim, Adrian Hutter, Priyanka Agrawal, Alex Castro-Ros, George van den Driessche, Tao Wang, Fan Yang, Shuo yiin Chang, Paul Komarek, Ross McIlroy, Mario Lučić, Guodong Zhang, Wael Farhan, Michael Sharman, Paul Natsev, Paul Michel, Yamini Bansal, Siyuan Qiao, Kris Cao, Siamak Shakeri, Christina Butterfield, Justin Chung, Paul Kishan Rubenstein, Shivani Agrawal, Arthur Mensch, Kedar Soparkar, Karel Lenc, Timothy Chung, Aedan Pope, Loren Maggiore, Jackie Kay, Priya Jhakra, Shibo Wang, Joshua Maynez, Mary Phuong, Taylor Tobin, Andrea Tacchetti, Maja Trebacz, Kevin Robinson, Yash Katariya, Sebastian Riedel, Paige Bailey, Kefan Xiao, Nimesh Ghelani, Lora Aroyo, Ambrose Slone, Neil Houlsby, Xuehan Xiong, Zhen Yang, Elena Gribovskaya, Jonas Adler, Mateo Wirth, Lisa Lee, Music Li, Thais Kagohara, Jay Pavagadhi, Sophie Bridgers, Anna Bortsova, Sanjay Ghemawat, Zafarali Ahmed, Tianqi Liu, Richard Powell, Vijay Bolina, Mariko Iinuma, Polina Zablotskaia, James Besley, Da-Woon Chung, Timothy Dozat, Ramona Comanescu, Xiance Si, Jeremy Greer, Guolong Su, Martin Polacek, Raphaël Lopez Kaufman, Simon Tokumine, Hexiang Hu, Elena Buchatskaya, Yingjie Miao, Mohamed Elhawaty, Aditya Siddhant, Nenad Tomasev, Jinwei Xing, Christina Greer, Helen Miller, Shereen Ashraf, Aurko Roy, Zizhao Zhang, Ada Ma, Angelos Filos, Milos Besta, Rory Blevins, Ted Klimenko, Chih-Kuan Yeh, Soravit Changpinyo, Jiaqi Mu, Oscar Chang, Mantas Pajarskas, Carrie Muir, Vered Cohen, Charline Le Lan, Krishna Haridasan, Amit Marathe, Steven Hansen, Sholto Douglas, Rajkumar Samuel, Mingqiu Wang, Sophia Austin, Chang Lan, Jiepu Jiang, Justin Chiu, Jaime Alonso Lorenzo, Lars Lowe Sjösund, Sébastien Cevey, Zach Gleicher, Thi Avrahami, Anudhyan Boral, Hansa Srinivasan, Vittorio Selo, Rhys May, Konstantinos Aisopos, Léonard Hussenot, Livio Baldini Soares, Kate Baumli, Michael B. Chang, Adrià Recasens, Ben Caine, Alexander Pritzel, Filip Pavetic, Fabio Pardo, Anita Gergely, Justin Frye, Vinay Ramasesh, Dan Horgan, Kartikeya Badola, Nora Kassner, Subhrajit Roy, Ethan Dyer, Víctor Campos Campos, Alex Tomala, Yunhao Tang, Dalia El Badawy, Elspeth White, Basil Mustafa, Oran Lang, Abhishek Jindal, Sharad Vikram, Zhitao Gong, Sergi Caelles, Ross Hemsley, Gregory Thornton, Fangxiaoyu Feng, Wojciech Stokowiec, Ce Zheng, Phoebe Thacker, Çağlar Ünlü, Zhishuai Zhang, Mohammad Saleh, James Svensson, Max Bileschi, Piyush Patil, Ankesh Anand, Roman Ring, Katerina Tsihlas, Arpi Vezer, Marco Selvi, Toby Shevlane, Mikel Rodriguez, Tom Kwiatkowski, Samira Daruki, Keran Rong, Allan Dafoe, Nicholas FitzGerald, Keren Gu-Lemberg, Mina Khan, Lisa Anne Hendricks, Marie Pellat, Vladimir Feinberg, James Cobon-Kerr, Tara Sainath, Maribeth Rauh, Sayed Hadi Hashemi, Richard Ives, Yana Hasson, Eric Noland, Yuan Cao, Nathan Byrd, Le Hou, Qingze Wang, Thibault Sottiaux, Michela Paganini, Jean-Baptiste Lespiau, Alexandre Moufarek, Samer Hassan, Kaushik Shivakumar, Joost van Amersfoort, Amol Mandhane, Pratik Joshi, Anirudh Goyal, Matthew Tung, Andrew Brock, Hannah Sheahan, Vedant Misra, Cheng Li, Nemanja Rakićević, Mostafa Dehghani, Fangyu Liu, Sid Mittal, Junhyuk Oh, Seb Noury, Eren Sezener, Fantine Huot, Matthew Lamm, Nicola De Cao, Charlie Chen, Sidharth Mudgal, Romina Stella, Kevin Brooks, Gautam Vasudevan, Chenxi Liu, Mainak Chain, Nivedita Melinkeri, Aaron Cohen, Venus Wang, Kristie Seymore, Sergey Zubkov, Rahul Goel, Summer Yue, Sai Krishnakumaran, Brian Albert, Nate Hurley, Motoki Sano, Anhad Mohananey, Jonah Joughin, Egor Filonov, Tomasz Kępa, Yomna Eldawy, Jiawern Lim, Rahul Rishi, Shirin Badiezadegan, Taylor Bos, Jerry Chang, Sanil Jain, Sri Gayatri Sundara Padmanabhan, Subha Puttagunta, Kalpesh Krishna, Leslie Baker, Norbert Kalb, Vamsi Bedapudi, Adam Kurzrok, Shuntong Lei, Anthony Yu, Oren Litvin, Xiang Zhou, Zhichun Wu, Sam Sobell, Andrea Siciliano, Alan Papir, Robby Neale, Jonas Bragagnolo, Tej Toor, Tina Chen, Valentin Anklin, Feiran Wang, Richie Feng, Milad Gholami, Kevin Ling, Lijuan Liu, Jules Walter, Hamid Moghaddam, Arun Kishore, Jakub Adamek, Tyler Mercado, Jonathan Mallinson, Siddhinita Wandekar, Stephen Cagle, Eran Ofek, Guillermo Garrido, Clemens Lombriser, Maksim Mukha, Botu Sun, Hafeezul Rahman Mohammad, Josip Matak, Yadi Qian, Vikas Peswani, Pawel Janus, Quan Yuan, Leif Schelin, Oana David, Ankur Garg, Yifan He, Oleksii Duzhyi, Anton Älgmyr, Timothée Lottaz, Qi Li, Vikas Yadav, Luyao Xu, Alex Chinien, Rakesh Shivanna, Aleksandr Chuklin, Josie Li, Carrie Spadine, Travis Wolfe, Kareem Mohamed, Subhabrata Das, Zihang Dai, Kyle He, Daniel von Dincklage, Shyam Upadhyay, Akanksha Maurya, Luyan Chi, Sebastian Krause, Khalid Salama, Pam G Rabinovitch, Pavan Kumar Reddy M, Aarush Selvan, Mikhail Dektiarev, Golnaz Ghiasi, Erdem Guven, Himanshu Gupta, Boyi Liu, Deepak Sharma, Idan Heimlich Shtacher, Shachi Paul, Oscar Akerlund, François-Xavier Aubet, Terry Huang, Chen Zhu, Eric Zhu, Elico Teixeira, Matthew Fritze, Francesco Bertolini, Liana-Eleonora Marinescu, Martin Bölle, Dominik Paulus, Khyatti Gupta, Tejasi Latkar, Max Chang, Jason Sanders, Roopa Wilson, Xuewei Wu, Yi-Xuan Tan, Lam Nguyen Thiet, Tulsee Doshi, Sid Lall, Swaroop Mishra, Wanming Chen, Thang Luong, Seth Benjamin, Jasmine Lee, Ewa Andrejczuk, Dominik Rabiej, Vipul Ranjan, Krzysztof Styrc, Pengcheng Yin, Jon Simon, Malcolm Rose Harriott, Mudit Bansal, Alexei Robsky, Geoff Bacon, David Greene, Daniil Mirylenka, Chen Zhou, Obaid Sarvana, Abhimanyu Goyal, Samuel Andermatt, Patrick Siegler, Ben Horn, Assaf Israel, Francesco Pongetti, Chih-Wei "Louis" Chen, Marco Selvatici, Pedro Silva, Kathie Wang, Jackson Tolins, Kelvin Guu, Roey Yogev, Xiaochen Cai, Alessandro Agostini, Maulik Shah, Hung Nguyen, Noah Ó Donnaile, Sébastien Pereira, Linda Friso, Adam Stambler, Adam Kurzrok, Chenkai Kuang, Yan Romanikhin, Mark Geller, ZJ Yan, Kane Jang, Cheng-Chun Lee, Wojciech Fica, Eric Malmi, Qijun Tan, Dan Banica, Daniel Balle, Ryan Pham, Yanping Huang, Diana Avram, Hongzhi Shi, Jasjot Singh, Chris Hidey, Niharika Ahuja, Pranab Saxena, Dan Dooley, Srividya Pranavi Potharaju, Eileen O'Neill, Anand Gokulchandran, Ryan Foley, Kai Zhao, Mike Dusenberry, Yuan Liu, Pulkit Mehta, Ragha Kotikalapudi, Chalence Safranek-Shrader, Andrew Goodman, Joshua Kessinger, Eran Globen, Prateek Kolhar, Chris Gorgolewski, Ali Ibrahim, Yang Song, Ali Eichenbaum, Thomas Brovelli, Sahitya Potluri, Preethi Lahoti, Cip Baetu, Ali Ghorbani, Charles Chen, Andy Crawford, Shalini Pal, Mukund Sridhar, Petru Gurita, Asier Mujika, Igor Petrovski, Pierre-Louis Cedoz, Chenmei Li, Shiyuan Chen, Niccolò Dal Santo, Siddharth Goyal, Jitesh Punjabi, Karthik Kappaganthu, Chester Kwak, Pallavi LV, Sarmishta Velury, Himadri Choudhury, Jamie Hall, Premal Shah, Ricardo Figueira, Matt Thomas, Minjie Lu, Ting Zhou, Chintu Kumar, Thomas Jurdi, Sharat Chikkerur, Yenai Ma, Adams Yu, Soo Kwak, Victor Ähdel, Sujeevan Rajayogam, Travis Choma, Fei Liu, Aditya Barua, Colin Ji, Ji Ho Park, Vincent Hellendoorn, Alex Bailey, Taylan Bilal, Huanjie Zhou, Mehrdad Khatir, Charles Sutton, Wojciech Rzadkowski, Fiona Macintosh, Konstantin Shagin, Paul Medina, Chen Liang, Jinjing Zhou, Pararth Shah, Yingying Bi, Attila Dankovics, Shipra Banga, Sabine Lehmann, Marissa Bredesen, Zifan Lin, John Eric Hoffmann, Jonathan Lai, Raynald Chung, Kai Yang, Nihal Balani, Arthur Bražinskas, Andrei Sozanschi, Matthew Hayes, Héctor Fernández Alcalde, Peter Makarov, Will Chen, Anto-

nio Stella, Liselotte Snijders, Michael Mandl, Ante Kärrman, Paweł Nowak, Xinyi Wu, Alex Dyck, Krishnan Vaidyanathan, Raghavender R, Jessica Mallet, Mitch Rudominer, Eric Johnston, Sushil Mittal, Akhil Udathu, Janara Christensen, Vishal Verma, Zach Irving, Andreas Santucci, Gamaleldin Elsayed, Elnaz Davoodi, Marin Georgiev, Ian Tenney, Nan Hua, Geoffrey Cideron, Edouard Leurent, Mahmoud Alnahlawi, Ionut Georgescu, Nan Wei, Ivy Zheng, Dylan Scandinaro, Heinrich Jiang, Jasper Snoek, Mukund Sundararajan, Xuezhi Wang, Zack Ontiveros, Itay Karo, Jeremy Cole, Vinu Rajashekhar, Lara Tumeh, Eyal Ben-David, Rishub Jain, Jonathan Uesato, Romina Datta, Oskar Bunyan, Shimu Wu, John Zhang, Piotr Stanczyk, Ye Zhang, David Steiner, Subhajit Naskar, Michael Azzam, Matthew Johnson, Adam Paszke, Chung-Cheng Chiu, Jaume Sanchez Elias, Afroz Mohiuddin, Faizan Muhammad, Jin Miao, Andrew Lee, Nino Vieillard, Jane Park, Jiageng Zhang, Jeff Stanway, Drew Garmon, Abhijit Karmarkar, Zhe Dong, Jong Lee, Aviral Kumar, Luowei Zhou, Jonathan Evens, William Isaac, Geoffrey Irving, Edward Loper, Michael Fink, Isha Arkatkar, Nanxin Chen, Izhak Shafran, Ivan Petrychenko, Zhe Chen, Johnson Jia, Anselm Levskaya, Zhenkai Zhu, Peter Grabowski, Yu Mao, Alberto Magni, Kaisheng Yao, Javier Snaider, Norman Casagrande, Evan Palmer, Paul Suganthan, Alfonso Castaño, Irene Giannoumis, Wooyeol Kim, Mikołaj Rybiński, Ashwin Sreevatsa, Jennifer Prendki, David Soergel, Adrian Goedeckemeyer, Willi Gierke, Mohsen Jafari, Meenu Gaba, Jeremy Wiesner, Diana Gage Wright, Yawen Wei, Harsha Vashisht, Yana Kulizhskaya, Jay Hoover, Maigo Le, Lu Li, Chimezie Iwuanyanwu, Lu Liu, Kevin Ramirez, Andrey Khorlin, Albert Cui, Tian LIN, Marcus Wu, Ricardo Aguilar, Keith Pallo, Abhishek Chakladar, Ginger Perng, Elena Allica Abellan, Mingyang Zhang, Ishita Dasgupta, Nate Kushman, Ivo Penchev, Alena Repina, Xihui Wu, Tom van der Weide, Priya Ponnapalli, Caroline Kaplan, Jiri Simsa, Shuangfeng Li, Olivier Dousse, Fan Yang, Jeff Piper, Nathan Ie, Rama Pasumarthi, Nathan Lintz, Anitha Vijayakumar, Daniel Andor, Pedro Valenzuela, Minnie Lui, Cosmin Paduraru, Daiyi Peng, Katherine Lee, Shuyuan Zhang, Somer Greene, Duc Dung Nguyen, Paula Kurylowicz, Cassidy Hardin, Lucas Dixon, Lili Janzer, Kiam Choo, Ziqiang Feng, Biao Zhang, Achintya Singhal, Dayou Du, Dan McKinnon, Natasha Antropova, Tolga Bolukbasi, Orgad Keller, David Reid, Daniel Finchelstein, Maria Abi Raad, Remi Crocker, Peter Hawkins, Robert Dadashi, Colin Gaffney, Ken Franko, Anna Bulanova, Rémi Leblond, Shirley Chung, Harry Askham, Luis C. Cobo, Kelvin Xu, Felix Fischer, Jun Xu, Christina Sorokin, Chris Alberti, Chu-Cheng Lin, Colin Evans, Alek Dimitriev, Hannah Forbes, Dylan Banarse, Zora Tung, Mark Omernick, Colton Bishop, Rachel Sterneck, Rohan Jain, Jiawei Xia, Ehsan Amid, Francesco Piccinno, Xingyu Wang, Praseem Banzal, Daniel J. Mankowitz, Alex Polozov, Victoria Krakovna, Sasha Brown, MohammadHossein Bateni, Dennis Duan, Vlad Firoiu, Meghana Thotakuri, Tom Natan, Matthieu Geist, Ser tan Girgin, Hui Li, Jiayu Ye, Ofir Roval, Reiko Tojo, Michael Kwong, James Lee-Thorp, Christopher Yew, Danila Sinopalnikov, Sabela Ramos, John Mellor, Abhishek Sharma, Kathy Wu, David Miller, Nicolas Sonnerat, Denis Vnukov, Rory Greig, Jennifer Beattie, Emily Caveness, Libin Bai, Julian Eisenschlos, Alex Korchemniy, Tomy Tsai, Mimi Jasarevic, Weize Kong, Phuong Dao, Zeyu Zheng, Frederick Liu, Fan Yang, Rui Zhu, Tian Huey Teh, Jason Sanmiya, Evgeny Gladchenko, Nejc Trdin, Daniel Toyama, Evan Rosen, Sasan Tavakkol, Linting Xue, Chen Elkind, Oliver Woodman, John Carpenter, George Papamakarios, Rupert Kemp, Sushant Kafle, Tanya Grunina, Rishika Sinha, Alice Talbert, Diane Wu, Denese Owusu-Afriyie, Cosmo Du, Chloe Thornton, Jordi Pont-Tuset, Pradyumna Narayana, Jing Li, Saaber Fatehi, John Wieting, Omar Ajmeri, Benigno Uria, Yeongil Ko, Laura Knight, Amélie Héliou, Ning Niu, Shane Gu, Chenxi Pang, Yeqing Li, Nir Levine, Ariel Stolovich, Rebeca Santamaria-Fernandez, Sonam Goenka, Wenny Yustalim, Robin Strudel, Ali Elqursh, Charlie Deck, Hyo Lee, Zonglin Li, Kyle Levin, Raphael Hoffmann, Dan Holtmann-Rice, Olivier Bachem, Sho Arora, Christy Koh, Soheil Hassas Yeganeh, Siim Põder, Mukarram Tariq, Yanhua Sun, Lucian Ionita, Mojtaba Seyedhosseini, Pouya Tafti, Zhiyu Liu, Anmol Gulati, Jasmine Liu, Xinyu Ye, Bart Chrzaszcz, Lily Wang, Nikhil Sethi, Tianrun Li, Ben Brown, Shreya Singh, Wei Fan, Aaron Parisi, Joe Stanton, Vinod Koverkathu, Christopher A. Choquette-Choo, Yunjie Li, TJ Lu, Abe Ittycheriah, Prakash Shroff, Mani Varadarajan, Sanaz Bahargam, Rob Willoughby, David Gaddy, Guillaume Desjardins, Marco Cornero, Brona Robenek, Bhavishya Mittal, Ben Albrecht, Ashish Shenoy, Fedor Moiseev, Henrik Jacobsson, Alireza Ghaffarkhah, Morgane Rivière, Alanna Walton, Clément Crepy, Alicia Parrish, Zongwei Zhou, Clement Farabet, Carey Radebaugh, Praveen Srinivasan, Claudia van der Salm, Andreas Fidjeland, Salvatore Scellato, Eri Latorre-Chimoto, Hanna Klimczak-Plucińska, David Bridson, Dario de Cesare, Tom Hudson, Piermaria Mendolicchio, Lexi Walker, Alex Morris, Matthew Mauger, Alexey Guseynov, Alison Reid, Seth Odoom, Lucia Loher, Victor Cotruta, Madhavi Yenugula, Dominik Grewe, Anastasia Petrushkina, Tom Duerig, Antonio Sanchez, Steve Yadlowsky, Amy Shen, Amir Globerson, Lynette Webb, Sahil Dua, Dong Li, Surya Bhupatiraju, Dan Hurt, Haroon Qureshi, Ananth Agarwal, Tomer Shani, Matan Eyal, Anuj Khare, Shreyas Rammohan Belle, Lei Wang, Chetan Tekur, Mihir Sanjay Kale, Jinliang Wei, Ruoxin Sang, Brennan Saeta, Tyler Liechty, Yi Sun, Yao Zhao, Stephan Lee, Pandu Nayak, Doug Fritz, Manish Reddy Vuyyuru, John Aslanides, Nidhi Vyas, Martin Wicke, Xiao Ma, Evgenii Eltyshev, Nina Martin, Hardie Cate, James Manyika, Keyvan Amiri, Yelin Kim, Xi Xiong, Kai Kang, Florian Luisier, Nilesh Tripuraneni, David Madras, Mandy Guo, Austin Waters, Oliver Wang, Joshua Ainslie, Jason Baldridge, Han Zhang, Garima Pruthi, Jakob Bauer, Feng Yang, Riham Mansour, Jason Gelman, Yang Xu, George Polovets, Ji Liu, Honglong Cai, Warren Chen, XiangHai Sheng, Emily Xue, Sherjil Ozair, Christof

Angermueller, Xiaowei Li, Anoop Sinha, Weiren Wang, Julia Wiesinger, Emmanouil Koukoumidis, Yuan Tian, Anand Iyer, Madhu Gurumurthy, Mark Goldenson, Parashar Shah, MK Blake, Hongkun Yu, Anthony Urbanowicz, Jennimaria Palomaki, Chrisantha Fernando, Ken Durden, Harsh Mehta, Nikola Momchev, Elahe Rahimtoroghi, Maria Georgaki, Amit Raul, Sebastian Ruder, Morgan Redshaw, Jinhyuk Lee, Denny Zhou, Komal Jalan, Dinghua Li, Blake Hechtman, Parker Schuh, Milad Nasr, Kieran Milan, Vladimir Mikulik, Juliana Franco, Tim Green, Nam Nguyen, Joe Kelley, Aroma Mahendru, Andrea Hu, Joshua Howland, Ben Vargas, Jeffrey Hui, Kshitij Bansal, Vikram Rao, Rakesh Ghiya, Emma Wang, Ke Ye, Jean Michel Sarr, Melanie Moranski Preston, Madeleine Elish, Steve Li, Aakash Kaku, Jigar Gupta, Ice Pasupat, Da-Cheng Juan, Milan Someswar, Tejvi M., Xinyun Chen, Aida Amini, Alex Fabrikant, Eric Chu, Xuanyi Dong, Amruta Muthal, Senaka Buthpitiya, Sarthak Jauhari, Nan Hua, Urvashi Khandelwal, Ayal Hitron, Jie Ren, Larissa Rinaldi, Shahar Drath, Avigail Dabush, Nan-Jiang Jiang, Harshal Godhia, Uli Sachs, Anthony Chen, Yicheng Fan, Hagai Taitelbaum, Hila Noga, Zhuyun Dai, James Wang, Chen Liang, Jenny Hamer, Chun-Sung Ferng, Chenel Elkind, Aviel Atias, Paulina Lee, Vít Listík, Mathias Carlen, Jan van de Kerkhof, Marcin Pikus, Krunoslav Zaher, Paul Müller, Sasha Zykova, Richard Stefanec, Vitaly Gatsko, Christoph Hirnschall, Ashwin Sethi, Xingyu Federico Xu, Chetan Ahuja, Beth Tsai, Anca Stefanoiu, Bo Feng, Keshav Dhandhania, Manish Katyal, Akshay Gupta, Atharva Parulekar, Divya Pitta, Jing Zhao, Vivaan Bhatia, Yashodha Bhavnani, Omar Alhadlaq, Xiaolin Li, Peter Danenberg, Dennis Tu, Alex Pine, Vera Filippova, Abhipso Ghosh, Ben Limonchik, Bhargava Urala, Chaitanya Krishna Lanka, Derik Clive, Yi Sun, Edward Li, Hao Wu, Kevin Hongtongsak, Ianna Li, Kalind Thakkar, Kuanysh Omarov, Kushal Majmundar, Michael Alverson, Michael Kucharski, Mohak Patel, Mudit Jain, Maksim Zabelin, Paolo Pelagatti, Rohan Kohli, Saurabh Kumar, Joseph Kim, Swetha Sankar, Vineet Shah, Lakshmi Ramachandruni, Xiangkai Zeng, Ben Bariach, Laura Weidinger, Tu Vu, Amar Subramanya, Sissie Hsiao, Demis Hassabis, Koray Kavukcuoglu, Adam Sadovsky, Quoc Le, Trevor Strohman, Yonghui Wu, Slav Petrov, Jeffrey Dean, and Oriol Vinyals. 2024. Gemini: A family of highly capable multimodal models. *Preprint*, arXiv:2312.11805.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Rowan Zellers, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. From recognition to cognition: Visual commonsense reasoning. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Gengyuan Zhang, Yurui Zhang, Kerui Zhang, and Volker Tresp. 2024. Can vision-language models be a good guesser? exploring vlms for times and location reasoning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 636–645.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.

A Additional Experiments

A.1 Experiment on Smaller Subset

One potential reason for the VLMs' consistently low performance on the TARA dataset could be its inherent difficulty, even for humans, in inferring the time and location from the images. To explore this, we manually selected 50 data points where the images were considered informative and indicative of time and location. We then conducted experiments on this subset. The results, shown in Table Table 7, demonstrate a significant performance improvement for our method, while BLIP2 and LLaVA did not show similar improvements. This suggests that although some data points in TARA are extremely challenging, the consistent marginal performance of BLIP2 and LLaVA indicates their inability to effectively handle the dataset's visual clues. In contrast, PuzzleGPT exhibited a notable improvement, highlighting its robustness and superior ability to utilize information from the image.

A.2 Perceiver Ablation

We ablated the performance of our BLIP-2 perceiver by replacing it with LLaVA. The results are shown in Table 8. Using BLiP2 as the perceiver outperformed using LLaVA, especially on location scores. This might be due to LLaVA's worse performance on location reasoning compared to BLiP2. For time Example-F1, using LLaVA as perceiver scored 43.46 with a brevity penalty, which is still better than the LLaVA baseline. This suggests using different backbones as the perceiver will inherently affect the models' output nature but generally elevate the performance compared to the backbone.

	Time		Location	
Model	**Acc(%)**	**Example-F1$^\beta$**	**Acc(%)**	**Example-F1$^\beta$**
BLiP2	4.88	38.59	*22.92*	38.55
LLaVA	*9.76*	*43.25*	*22.92*	*40.34*
PuzzleGPT	**12.20**	**46.24**	**35.42**	**57.75**

Table 7: Experiment conducted on a **smaller (50)** subset that are manually selected by human evaluator. Instances in this dataset are considered information rich, while generative VLMs failed to receive performance improvement.

	Time		Location	
Model	**Acc(%)**	**Example-F1$^\beta$**	**Acc*(%)**	**Example-F1$^\beta$**
BLiP2	0.30	32.27	*17.41*	*43.59*
LLaVA	0.23	43.26	7.85	25.92
PuzzleGPT	0.30	**43.72**	**22.99**	**51.04**
PuzzleGPT(LLaVA)	0.30	*43.46*	13.71	31.92

Table 8: Performance drops when switching from BLiP2 to LLaVA in PuzzleGPT. We discovered a significant drop on location performance, which is consistent to the location performance gap between BLiP2 and LLaVA. With a stronger perceiver, a better performance might be expected.

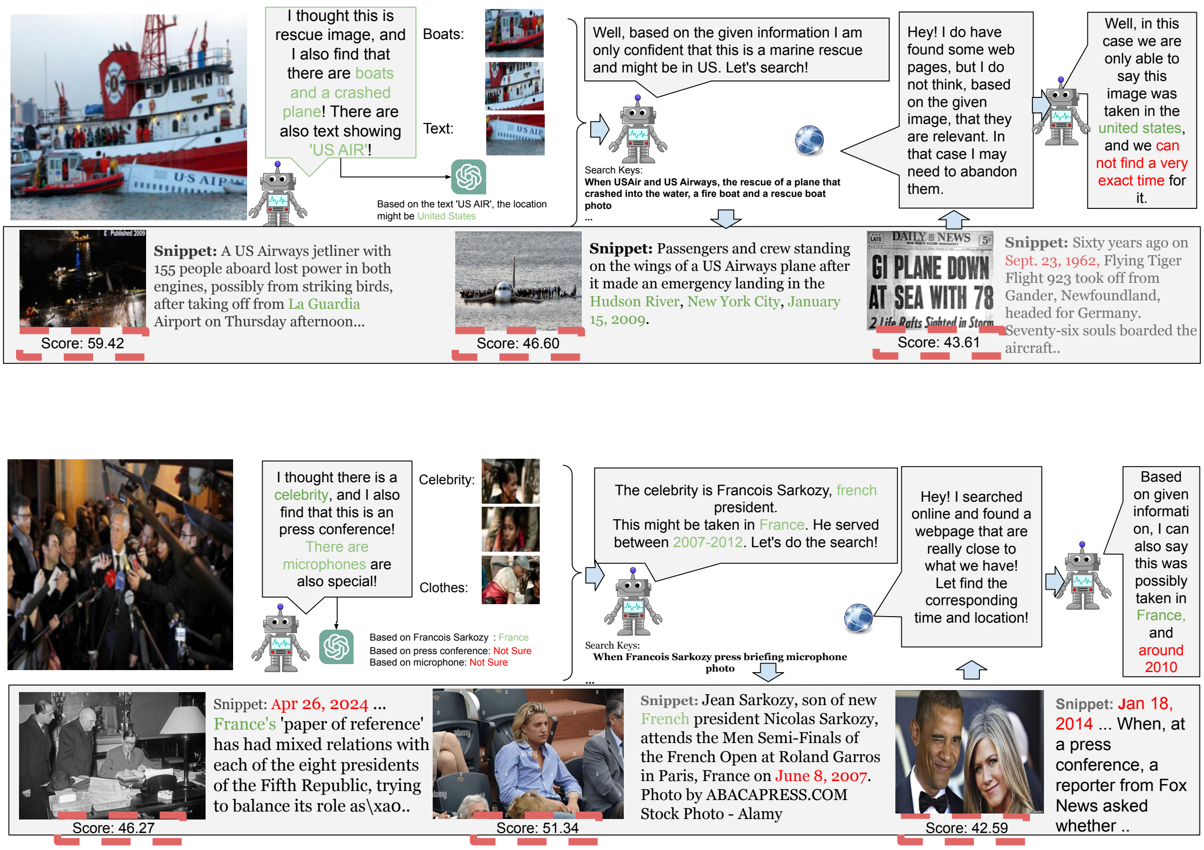

Figure 10: Two samples for negative case studies.

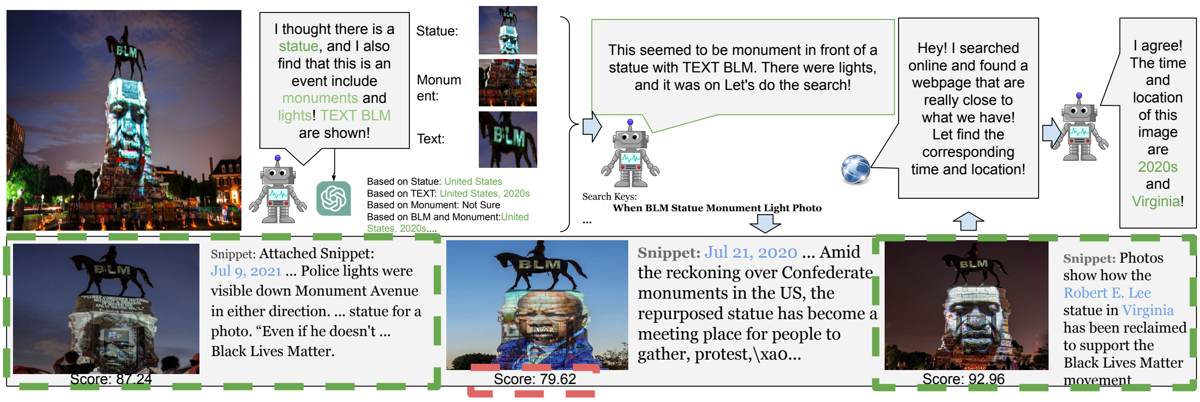

Figure 11: Another positive case study.