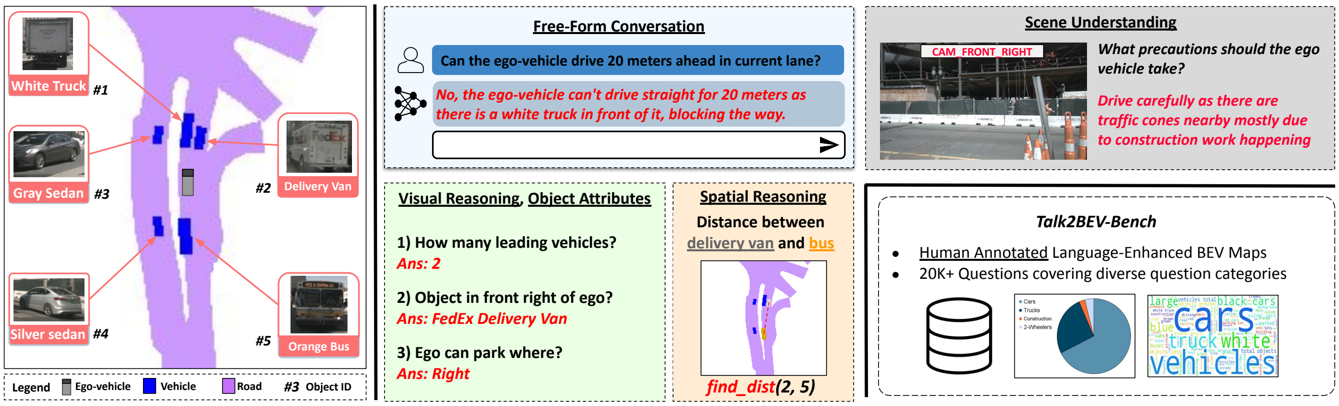
Talk2BEV: Language-enhanced Bird's-eye View Maps for Autonomous Driving

https://llmbev.github.io/talk2bev/

Vikrant Dewangan*[1], Tushar Choudhary*[1], Shivam Chandhok*[2], Shubham Priyadarshan[1], Anushka Jain[1], Arun K. Singh[3], Siddharth Srivastava[4], Krishna Murthy Jatavallabhula[5], and K. Madhava Krishna[1]
[1]IIIT Hyderabad, [2]University of British Columbia, [3]University of Tartu, [4]TensorTour Inc., [5]MIT



Free-Form Conversation

Can the ego-vehicle drive 20 meters ahead in current lane?

No, the ego-vehicle can't drive straight for 20 meters as there is a white truck in front of it, blocking the way.

Scene Understanding

What precautions should the ego vehicle take?

Drive carefully as there are traffic cones nearby mostly due to construction work happening

Visual Reasoning, Object Attributes

1) How many leading vehicles?
Ans: 2

2) Object in front right of ego?
Ans: FedEx Delivery Van

3) Ego can park where?
Ans: Right

Spatial Reasoning

Distance between delivery van and bus

find_dist(2, 5)

Talk2BEV-Bench

- Human Annotated Language-Enhanced BEV Maps
- 20K+ Questions covering diverse question categories

Fig. 1: ***Talk2BEV*** builds *Language-enhanced Bird's-Eye View (BEV) maps* using (a) BEV representations constructed from vehicle sensors (multi-view images, lidar), and (b) Aligned vision-language features for each object which can be directly used as context within large vision-language models (LVLMs) to query and *talk* to the objects in the scene. These maps embed knowledge about object semantics, material properties, affordances, and spatial concepts and can be queried for visual reasoning, spatial understanding, and making decisions about potential future scenarios, critical for autonomous driving application. Further, we develop the first benchmark *Talk2BEV-Bench* towards evaluate LVLMs for AD applications spanning a diverse set of question categories with human-annotated ground-truth.

***Abstract*— This work introduces *Talk2BEV*, a large vision-language model (LVLM)[1] interface for bird's-eye view (BEV) maps commonly used in autonomous driving. While existing perception systems for autonomous driving scenarios have largely focused on a pre-defined (closed) set of object categories and driving scenarios, *Talk2BEV* eliminates the need for BEV-specific training, relying instead on performant pre-trained LVLMs. This enables a single system to cater to a variety of autonomous driving tasks encompassing visual and spatial reasoning, predicting the intents of traffic actors, and decision-making based on visual cues. We extensively evaluate *Talk2BEV* on a large number of scene understanding tasks that rely on *both* the ability to interpret freefrom natural language queries, and in grounding these queries to the visual context embedded into the language-enhanced BEV map. To enable further research in LVLMs for autonomous driving scenarios, we develop and release *Talk2BEV-Bench*, a benchmark encompassing 1000 human-annotated BEV scenarios, with more than 20,000 questions and ground-truth responses from the NuScenes dataset. We encourage the reader to view the demos on our project page: https://llmbev.github.io/talk2bev/**

I. Introduction

For safe navigation without human intervention, autonomous driving (AD) systems need to understand the visual world around them to make informed decisions. This entails not just recognizing specific object categories, but also contextualizing their current and potential future interactions with the environment. Existing AD systems rely on domain-specific models for each scene understanding task, such as detecting traffic actors and signage or forecasting plausible future events. On the other hand, recent advances in large language models (LLMs) [4]–[8] and large vision-language models (LVLMs) [2], [3], [9], [10] have demonstrated a promising alternative to thinking about perception for AD; that of a single model pretrained on web-scale data, capable of performing all the aforementioned tasks and more (particularly, the ability to deal with unforeseen scenarios). In this work we ask, *how do we most efficiently integrate such capabilities of LLMs with scene represenations traditionally used in autonomous driving*?

To this end, we introduce *Talk2BEV*, language-enhanced maps for AD that enable holistic scene understanding and reasoning across a broad class of road scenarios. Our framework interfaces LVLMs with bird's-eye view (BEV) maps—top-down semantic maps of the road plane and traffic actors that are widely used in AD systems [11]–[14] —to enable visual reasoning, spatial understanding, and decision-making. Specifically, our model takes in a BEV map of a scene and augments the BEV with aligned image-language features for each object in the scene. These features can directly be passed as context to an LVLM, enabling the model to answer a wide range of questions about the scene and make decisions about potential future scenarios using the vast knowledge base acquired by the LVLM during training.

*Equal contribution.

[1] In this work, we use this term to refer to instruction-finetuned vision-language models; i.e., models that can consume text and image as input, and outupt text conditioned on the visual context [1]–[3].

We find that these LVLMs can interpret object semantics, material properties, affordances, and spatial concepts; and are an ideal alternative to domain-specific models. Notably, our approach does not require any BEV-specific or vision-language training/finetuning; and uses existing pretrained LLMs and LVLMs. This allows our approach to be flexibly and raplidly deployed on a wide class of domains and tasks, and to easily adapt to newer LLMs and LVLMs as newer and better models become available.

To objectively evaluate LVLMs for perception in the AD context and to expedite further reserch, we also develop Talk2BEV-Bench: a benchmark that assesses autonomous driving systems on both scene-level and object-level understanding. In summary, our contributions are as follows

- We develop *Talk2BEV*, the first system to augment BEV maps with language to enable general-purpose visuolinguistic reasoning for AD scenarios.
- Our framework does not require any training or finetuning, relying instead on pre-trained image-language models. This allows generalization to a diverse collection of models, scenarios, and tasks.
- We develop Talk2BEV-Bench, a large-scale benchmark for evaluating LVLMs for AD applications, including, but not limited to object attributes, semantics, visual reasoning, spatial understanding, and decision-making.

II. Related work

Large Vision Language Models. Following the rapid adoption of large language models (LLM) [4]–[8], several large vision-language models (LVLMs) [2], [3], [9], [10] have been released over the last few months. These models are tailored for visuolingustic tasks and are trained on pairs of aligned images and text. Evaluating and benchmarking these models, however, remains a challenge. Benchmarks [15]–[18] have gradually developed objective techniques to assess LVLMs, where a common theme is the curation of question-response pairs using off-the-shelf LLMs, and the responses are expected to one of the multiple choices provided.

3D Vision-Language Models. Progress in 3D scene comprehension has been significantly aided by language models such as in 3D object localization [19]–[21] and captioning [22], [23]. 3D Visual Question Answering, utilizing multi-view images [24], [25] or point clouds [26], [27], addresses object spatial reasoning and geometry. 3D-LLM [28] integrates LLMs into point clouds from multi-view images, bridging 2D models to 3D. In contrast, Point-LLM [29] trains solely on point clouds, bypassing images.

Multimodal Tasks in Autonomous Driving. Language-guided navigation and objeect-referral are recently being explored into autonomous driving [30]–[32]. Approaches such as CityScapes-Ref [33], Talk2Car [31] attempt object-referral on CityScapes [34] and NuScenes [35], respectively. ReferKITTI [36] merges temporal data for object referral and Multi-Object Tracking (MOT) on the KITTI dataset, while NuPrompt [32] takes a 3D approach with RoBERTa [37] as their language encoder. For language-based scene-understanding, there exist few works. NuScenes-QA [38] addresses Visual Question Answering (VQA) in autonomous driving by crafting scene graphs and question templates. Their evaluation demands end-to-end training and exact answer matching. We adopt Bird's Eye View (BEV) representations from a BEV network [11], answering open-ended queries using a 2D multimodal map powered by LVLM and GPT-4 for fairer evaluations. In contrast to earlier methods, we offer zero-shot scene comprehension using LVLM's generalization and introde a broader benchmark, *Talk2BEV-Bench*, to assess LVLMs for scene understanding via BEVs in autonomous driving.

III. Talk2BEV

The key idea of *Talk2BEV* is to embed a birds-eye view (BEV) map with general-purpose vision-language features derived from pretrained LVLMs. A BEV map, denoted $\mathcal{O}$, is a top-view multi-channel grid encoding semantic information like vehicle, road, lanes, etc. The ego-vehicle is at the origin, assumed to be the center of the BEV. Given multi-view RGB images $\mathcal{I}$ a LiDAR pointcloud $\mathcal{X}$, a BEV can be obtained using a number of off-the-shelf approaches [11], [12], [14], [39], [40].

Our three-phase pipeline (see Fig. 2) proceeds as follows:

1) We first estimate a BEV map using onboard vehicle sensors (multi-view images) using an off-the-shelf BEV prediction model [11].
2) For each object in this BEV map, we generate aligned image-language features using an LVLM [1], [2], [10]. These features are then passed into the language model of an LVLM to extract object metadata. The object data, in conjunction with geometric information encapsulated in the BEV, forms the language-enhanced map, $\mathbf{L}(\mathcal{O})$.
3) Finally, given a user query, we prompt an LLM (eg. GPT-4 [9]) which interprets this query, parses the language-enhanced BEV as needed, and produces a response to this query.

A. Language Enhanced Maps

BEV-Image Correspondence. First, we localize each object in the estimated BEV across the multi-view images used to produce the BEV map. For each object in the BEV map, we compute a set of k closest points in the lidar scan (a pointcloud); and project them into the camera frame using an inverse homography.

Map Representation. Our language-enhanced map augments the set of objects in a BEV by computing the image region corresponding to the object and deriving spatial and textual descriptions. For each object i, we compute (a) displacement along the BEV X and Y axes (in m) from the ego-vehicle, (b) object area (in m^2), (c) a text description of the object, and (d) a text description of the background. LVLMs are specifically prompted to generated detailed descriptions of objects, and their outputs typically encode the type, color, and utility of the vehicle, status of the vehicle indicators, any text displayed on the vehicle, and more[2].

[2] All prompts we used are made available on our webpage.

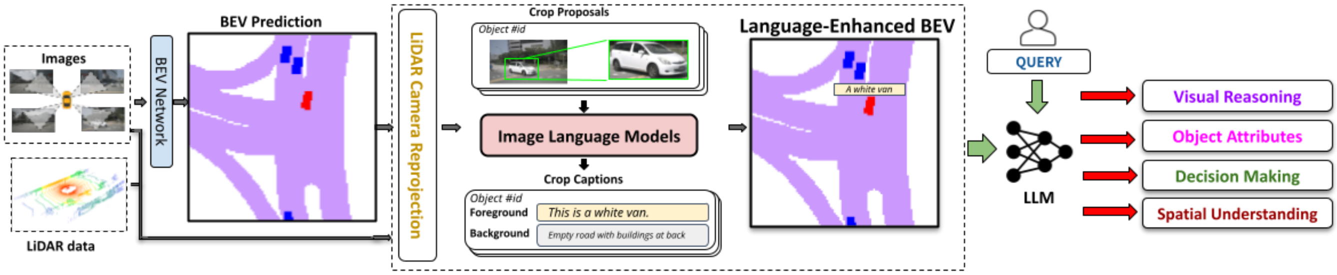


Fig. 2: **Overall *Talk2BEV* Pipeline**: We first generate Bird's Eye View (BEV) map $\mathcal{O}$ from the perspective images $\mathcal{I}$ and Lidar data. Next we construct the Language-Enhanced map $\mathbf{L}(\mathcal{O})$ by augmenting the generated BEV with aligned image-language features for each object from large vision-language models (LVLMs). These features can directly be used as context to LVLMs for answering object and scene-specific queries. The comprehensive Language Enhanced Map representation encodes objects in the scene along with their semantic descriptions and geometric cues from BEV . For each object i in the BEV Prediction $\mathcal{O}$, we extract its crop proposals r_i from the LiDAR-Camera re-projection pipeline, and obtain its captions using Large Vision-Language Models (LVLMs). Each object i contains in the map $\mathbf{L}$ its bev information including geometric cues like - area, centroid, and object descriptions like crop captions.

Language-enhancement. We then use the point-prompt feature of FastSAM [41] to generate masks specific to the region of the BEV crop. The back-projected points serve as positive labels to the point-prompt. For each region output of FastSAM, we obtain the bounding box b_i around it. The cropped image is then passed onto the LVLM, to generate crop descriptions. Specifically, we pass the crop to LVLM visual branch, to obtain the aligned image-language features. These features is directly passed as context in form of initial tokens to the language decoder branch of LVLM (say, Vicuna) to decode them. The descriptions for each object encompass both instance-level and scene-level details.

B. Response Generation

Type of queries. The *Talk2BEV* system can handle three types of user queries: Free-form q_{ff}, Multiple Choice Questions (MCQ) with one correct answer q_{mcq}, and Spatial Reasoning $q_{sp.}$, the query $q \in \{q_{mcq}, q_{sp.}, q_{ff}\}$. The Free-form q_{ff} and Spatial Reasoning $q_{sp.}$ queries allow us to assess general reasoning and decision making capabilities of our model. For a more comprehensive quantative evaluation, we include diverse queries from the the query types q_{mcq} and q_{sp} in our evaluation bench.

Spatial Operators We implement a set of primitive spatial operators [44] that parse complex natural language spatial reasoning queries q_{sp} into a set of function calls on the entities of the language-enhanced map $\mathbf{L}(\mathcal{O})$. These modules principally accept the object_id of the referenced objects and, when appropriate, the distance (m) as input. A comprehensive list of these spatial operators is provided in Table I. Depending on their return type, they can be categorized into two primary types (i) return list (comprising of object ids), and (ii) return distance. Their metrics are tailored based on this return type (details are in Sec. IV-B). This systematic approach facilitates the evaluation and scoring of these LVLMs, establishing a consistent format for assessment. Fig. 4 denotes an example usage of spatial operators. Notice that we are able to capture the distance between the construction vehicle and the truck carrying materials - 2 vehicles visible in different cameras. The Language Enhanced Maps for the objects are interpreted by an LLM to invoke relevant spatial operators in our framework to find the final distance.

C. Implementation Details

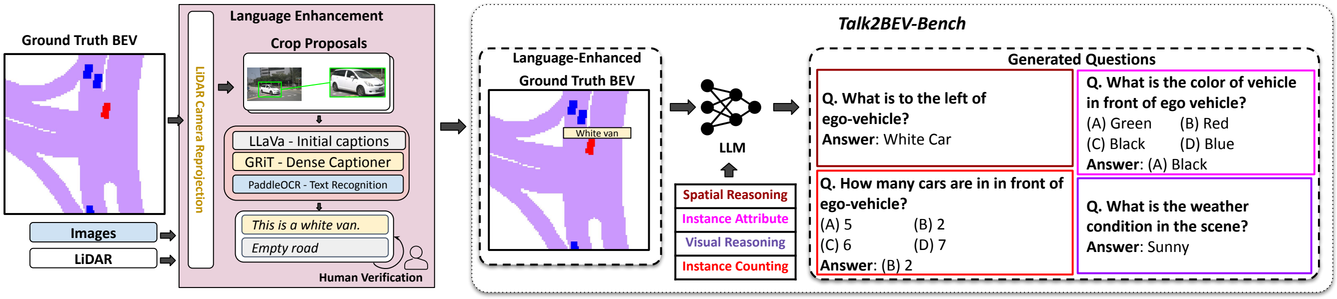
The BEV maps are generated from Lift-Splat-Shoot network [11] with dimension 200 × 200 pixels and are sampled with resolution of 0.5m, with the Ground Truth BEV having the same dimension and resolution. We use one of the LVLM models(i.e BLIP-2, MiniGPT-4 and InstructBLIP-2) to augment each object in our Language Enhanced Maps $\mathbf{L}(\mathcal{O})$ with corresponding visual features. These features are later used as context to language decoder of LVLM to output object descriptions. For BLIP-2, we use Flan5XXL [45] language decoder and for InstructBLIP-2/MiniGPT-4, we use the Vicuna-13b language decoder [46]. We use the default temperature value of 0.7 for LVLM for all experiments. We perform inference on NVIDIA DGX A100.

Method	**Description**
`front_filter(objs)`	objects to the front
`left_filter(objs)`	objects to the left
`right_filter(objs)`	objects to the right
`rear_filter(objs)`	objects to the rear
`dist_filter(objs, X)`	objects within **"X"**m
`k_closest(objs, k)`	k closest objects
`k_farthest(objs, k)`	k farthest objects
`objs_in_dist(objs, id, dist)`	objects within distance **"dist"** to o_{id}
`k_closest_to_obj(objs, id, k)`	k closest objects to o_{id}
`k_farthest_to_obj(objs, id, k)`	k farthest objects to o_{id}
`obj_distance(objs, id)`	distance (in m) to o_{id}
`find_dist(objs, id1, id2)`	distance between 2 objects o_{id1}, o_{id2}

TABLE I: **List of Spatial Operators used in *Talk2BEV*** Here *objs* refers to the list of objects in the BEV, o_{id} refers to an object with object_id as *id*. Note that wherever the object_id is not mentioned as inputs, the operator is applied on the ego-vehicle.

IV. Evaluation - Talk2BEV Bench

To evaluate the quality of our language-enhanced map $\mathbf{L}(\mathcal{O})$ and assess the spatial understanding and visual rea-



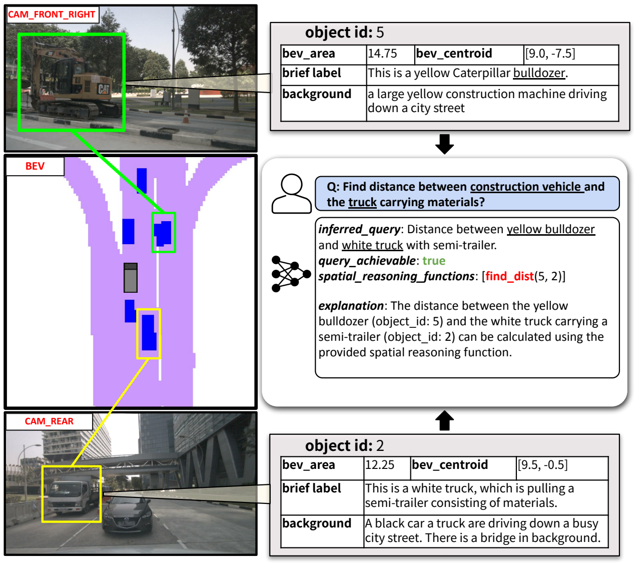
Q. What is to the left of ego-vehicle?
Answer: White Car

Q. What is the color of vehicle in front of ego vehicle?
(A) Green (B) Red
(C) Black (D) Blue
Answer: (A) Black

Q. How many cars are in in front of ego-vehicle?
(A) 5 (B) 2
(C) 6 (D) 7
Answer: (B) 2

Q. What is the weather condition in the scene?
Answer: Sunny

Fig. 3: *Talk2BEV-Bench* Creation: To develop this benchmark, we use the NuScenes Ground Truth BEV annotations and generate object and scene-level descriptions using dense Captioners (GRiT [42]), and Text-Recognition (PaddleOCR [43]) models. The Ground Truth BEV is then passed to an LLM like GPT4 to generate diverse questions including, but not limited to- Spatial Reasoning, Instance Attribute, Visual Reasoning and Instance Counting.



object id: 5

bev_area	14.75	**bev_centroid**	[9.0, -7.5]
brief label	This is a yellow Caterpillar bulldozer.		
background	a large yellow construction machine driving down a city street		

Q: Find distance between construction vehicle and the truck carrying materials?

inferred_query: Distance between yellow bulldozer and white truck with semi-trailer.
query_achievable: true
spatial_reasoning_functions: [find_dist(5, 2)]

explanation: The distance between the yellow bulldozer (object_id: 5) and the white truck carrying a semi-trailer (object_id: 2) can be calculated using the provided spatial reasoning function.

object id: 2

bev_area	12.25	**bev_centroid**	[9.5, -0.5]
brief label	This is a white truck, which is pulling a semi-trailer consisting of materials.		
background	A black car a truck are driving down a busy city street. There is a bridge in background.		

Fig. 4: **Spatial Operators**: To compute distance between bulldozer and white truck, the Language Enhanced Maps for the objects are interpreted by an LLM like GPT4 to invoke relevant spatial operators in our framework with appropriate object IDs as arguments.

soning capabilities of our framework, we present the first benchmark for assessing image-language models for autonomous driving applications, i.e, *Talk2BEV-Bench*. The Bench consisting of 1000 scenes, with their corresponding Language-Enhanced Ground Truth Maps $\mathbf{L}(\overline{\mathcal{O}})$, and more than 20k question-answer pairs [18] generated on these to assess and evaluate diverse aspects integral for autonomous driving. The benchmark contains queries q related to MCQs on object attributes, instance counting, visual reasoning, decision making and spatial reasoning (i.e q_{mcq} and q_{sp}). To develop this benchmark, we use the NuScenes Ground Truth BEV annotations and, taking inspiration from SEED-Bench [18], we use GPT-4 [9] with custom prompts to generate question-answer pairs. Fig. 3 shows the overview of the construction pipeline of our *Talk2BEV-Bench*. Next, we detail construction of Language-Enchanced Ground Truth Maps, Question Generation and the Evaluation Metrics.

A. Language-Enchanced Ground Truth Maps $\mathbf{L}(\overline{\mathcal{O}})$

To create the Ground Truth Maps $\mathbf{L}(\overline{\mathcal{O}})$, we use the BEV annotations $\overline{\mathcal{O}}$ from the NuScenes data to identify object and region proposals. For each object o_i and its associated region r_i, we extract its image-language description as follows.

Crop captions. We employ GRiT [42] for dense captioning to capture fine-grained details within the local crop. We also leverage PaddleOCR [43] for text recognition, extracting text from numerous vehicles to enhance understanding of their type and category, thereby improving Bench quality.

Background information: Beyond crop-level features, we extract vital contextual information crucial for autonomous driving application by extracting captions for the multi-view images. This includes street signs, barriers, weather conditions, time of day, and unique scene elements. Human annotators verify and refine the combined foreground and background captions at this stage, as shown in Fig. 3.

B. Question Generation and Evaluation Metrics

We cover diverse aspects in our benchmark, including object-level details like object attributes, counting as well as visual reasoning, decision making and spatial reasoning. For each scene and evaluation dimension, we prompt GPT-4 five times to generate five such questions per dimension, resulting in 20 questions per scene. The benchmark comprises two primary question types: MCQ - q_{mcq} and Spatial Reasoning - q_{sp}. For q_{mcq}, the bench contains its correct option. For q_{sp} which returns (*list of objects*), the bench contains list of objects extracted from the ground truth BEV with the relevant spatial operators. On the other hand, for q_{sp} of return type (*distance*), it returns the precise distance of the query.

For mcqs q_{mcq}, the generated response is evaluated against the correct option from the bench to obtain the accuracy.

For spatial queries q_{sp}, the response can be in two formats - *list of objects* or *distances* depending on the nature of the query. On return type is list, we use Intersection-Over-Union (IoU) metric to compare answer with ground-truth list, and in case of distances we use Distance Error. It evaluates the absolute difference between the distance from Bench and response and normalizes with the actual distance.

V. Results

We evaluate our framework *Talk2BEV* on queries which to cover diverse aspects relevant for Autonomous Driving applications. The queries include MCQs $\{q_{mcq}\}$ of type Visual Reasoning, Object Attributes, and Decision Making, q_{sp} - Spatial Reasoning. Each q_{mcq} comes from our *Talk2BEV-Bench* and allow quantitative evaluation of our framework. On the other hand, q_{sp} and q_{ff} allow qualitative evaluation of our framework. Through this comprehensive evaluation process - qualitative and quantitative, we seek to understand the efficiacy of our framework for driving application and evaluate the quality of our generated Language-Enhanced maps to asses their ability to answer crucial user queries. We comprehensively evaluate our performance using different LVLM models (i.e BLIP-2, InstructBLIP-2, MiniGPT-4) for constructing $\mathbf{L}(\mathcal{O})$. The LSS column refers to the case where BEV predicted using LSS method is used and GT refers to the case where ground-truth BEV is used.

A. Quantitative Results

Table II shows the performance of our framework on diverse questions from the Talk2BEV benchmark. For the LSS BEV, the $\mathbf{L}(\mathcal{O})$ constructed using InstructBLIP-2 achieves the best performance in instance attribute recognition and visual reasoning compared to the BLIP-2 and MiniGPT-4 counterparts. In contrast, for instance counting, MiniGPT-4 based $\mathbf{L}(\mathcal{O})$ map achieves the best accuracy. Overall, we notice that MiniGPT-4 achieves best average performance across different types of questions. Thanks to the conversation prompts dataset MiniGPT-4 uses which allow it to generate coherent text to answer user's questions and improve its usability [2]. We notice that instance attribute and visual reasoning tasks are more sensitive to the quality of LVLM captions compared to other types of question categories which is understandable given the complexity of these tasks compared to counting instances. As expected, the Language-Enhanced Maps based on ground-truth perform sligtly better than those constructed using LSS predicted BEVs. The small difference in performance here demonstrates that our LSS predicted BEVs are a good approximation of the ground-truth BEV maps.

BEV	LVLM	Instance Attribute	Instance Counting	Visual Reasoning	Avg
LSS	BLIP-2	0.50	0.83	0.47	0.60
	InstructBLIP-2	**0.54**	0.80	**0.50**	0.62
	MiniGPT-4	0.50	**0.90**	0.49	**0.63**
GT	BLIP-2	0.51	0.83	0.47	0.60
	InstructBLIP-2	**0.55**	0.80	0.50	0.62
	MiniGPT-4	**0.55**	**0.91**	**0.51**	**0.66**

TABLE II: **Overall Accuracy on MCQ Queries** (q_{mcq}). Performance of *Talk2BEV* with Language Enhanced Map constructed with different LVLMs (BLIP-2, InstructBLIP-2, MiniGPT-4) and BEV variants (LSS and GT) on Multiple Choice Questions (MCQs).

B. Qualitative Results

Fig. 5 showcases a scene with MCQ queries q_{mcq} with Language-enhanced maps - $\mathbf{L}_{BLIP2}$, $\mathbf{L}_{MiniGPT4}$, $\mathbf{L}_{InstructBLIP2}$ constructed using different LVLMs : BLIP-2 [1], InstructBLIP-2 [10], and MiniGPT-4 [2]. We highlight two BEV objects – a police car labeled 'Police' parked in front of the ego-vehicle and an orange construction truck with a crane located on the rear right of ego-vehicle. Corresponding *Talk2BEV-Bench* questions for these vehicles are displayed. We notice that map constructed with BLIP-2 identifies both the objects as white truck, leading to incorrect answers to *Talk2BEV-Bench* questions. In contrast, maps constructed with both MiniGPT-4 and InstructBLIP-2 identify the foreground object correctly leading to comparatively more correct answers than BLIP-2 variant. This indicates that the language enhanced map encoding the object attribute especially for those of the object crop is critical towards the overall performance. For crane, the detail of the foreground from InstructBLIP-2 based map is more specific i.e. 'Orange Crane with cab' than 'Large Orange Crane' from MiniGPT-4 variant. This distinction also leads to InstructBLIP-2 variant answering a question correctly under 'Instance Counting' about the count of white cars while other two models provide an incorrect answer. This states that our proposed workflow is successfully able to leverage the aligned-visual features obtained from LVLMs, and that the features embed enough object level semantics within Language enhanced map to correctly answer questions.

Fig. 6 illustrates a free-form interactive dialogue with the LLM where the user intends to advance by 20m and inquires about potential obstructions. Ahead of the ego-vehicle is a vehicle reversing into a spot. The Language-Enhanced map, $\mathbf{L}$, indicates its reverse light status and BEV position. Leveraging this, the LLM deduces the vehicle's intent and advises caution. The LLM's prediction aligns with the vehicle's future activity from $t = 0$ to $t = 3$s.

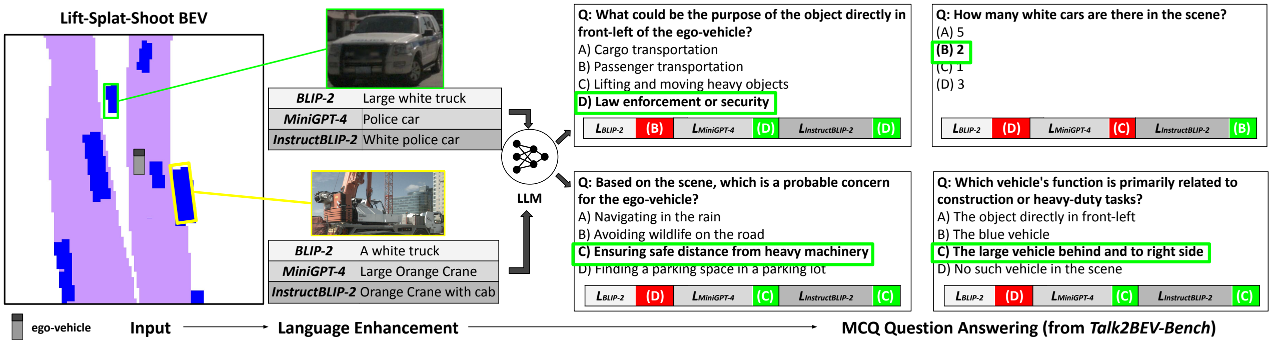
C. Impact of Spatial Operators.

	IoU ↑	Distance Error ↓
Random	0.16	0.44
Talk2BEV w/o SO*	0.25	0.22
Talk2BEV with SO*	0.83	0.13

*SO: Spatial Operators

TABLE III: **Evaluating Spatial Queries** q_{sp} - with IoU and Distance Error metric

Table III compares different techniques for spatial reasoning. Note that spatial reasoning queries are evaluated using IOU or distance error based on nature of query as explained in section IV. The *Random* method, which refers to randomly guessing distances and relevant objects, exhibits the lowest performance in terms of IoU score and the highest Distance Error. In comparison, Talk2BEV without Spatial Operators demonstrates relatively better performance when compared with the random baseline. Notice that Talk2BEV integrated with our Spatial Operators achieves superior performace in



BLIP-2	Large white truck
MiniGPT-4	Police car
InstructBLIP-2	White police car

BLIP-2	A white truck
MiniGPT-4	Large Orange Crane
InstructBLIP-2	Orange Crane with cab

Q: What could be the purpose of the object directly in front-left of the ego-vehicle?
A) Cargo transportation
B) Passenger transportation
C) Lifting and moving heavy objects
D) Law enforcement or security

*L*BLIP-2 (B) *L*MiniGPT-4 (D) *L*InstructBLIP-2 (D)

Q: How many white cars are there in the scene?
(A) 5
(B) 2
(C) 1
(D) 3

*L*BLIP-2 (D) *L*MiniGPT-4 (C) *L*InstructBLIP-2 (B)

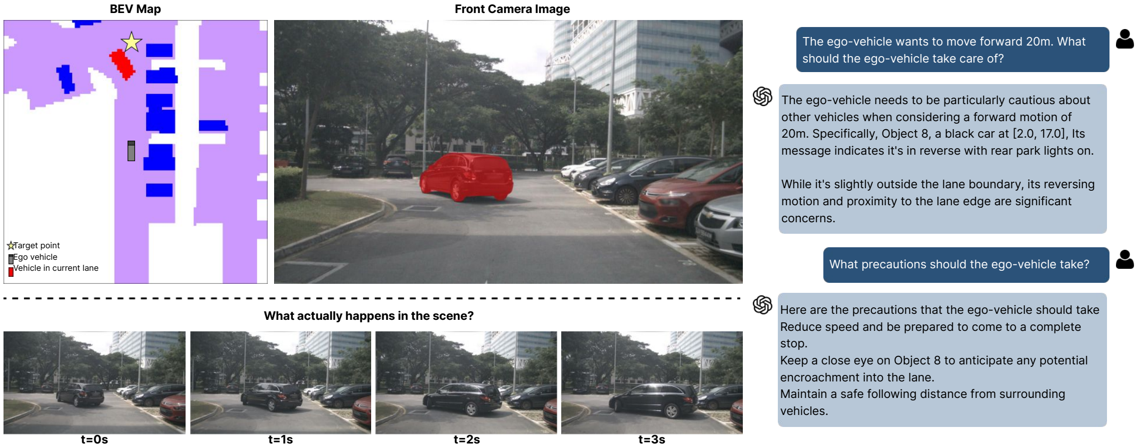
Q: Based on the scene, which is a probable concern for the ego-vehicle?
A) Navigating in the rain
B) Avoiding wildlife on the road
C) Ensuring safe distance from heavy machinery
D) Finding a parking space in a parking lot

*L*BLIP-2 (D) *L*MiniGPT-4 (C) *L*InstructBLIP-2 (C)

Q: Which vehicle's function is primarily related to construction or heavy-duty tasks?
A) The object directly in front-left
B) The blue vehicle
C) The large vehicle behind and to right side
D) No such vehicle in the scene

*L*BLIP-2 (D) *L*MiniGPT-4 (C) *L*InstructBLIP-2 (C)

Fig. 5: **Qualitative Results:** A BEV corresponding to a scene with multiple vehicles at an interchange. *Talk2BEV* is able to identify emergency vehicles (such as the *police car* shown here). The captions for a police car and a construction vehicle from Language Enhanced maps constructed with with different LVLMs (BLIP-2, InstructBLIP-2, MiniGPT-4) have been visualized. We show the corresponding BEV captions produced by various LVLMs and their performance across 4 questions from *Talk2BEV-Bench* relevant to these 2 objects. The correct answer for each question is highlighted in green.



The ego-vehicle wants to move forward 20m. What should the ego-vehicle take care of?

The ego-vehicle needs to be particularly cautious about other vehicles when considering a forward motion of 20m. Specifically, Object 8, a black car at [2.0, 17.0], Its message indicates it's in reverse with rear park lights on.

While it's slightly outside the lane boundary, its reversing motion and proximity to the lane edge are significant concerns.

What precautions should the ego-vehicle take?

Here are the precautions that the ego-vehicle should take
Reduce speed and be prepared to come to a complete stop.
Keep a close eye on Object 8 to anticipate any potential encroachment into the lane.
Maintain a safe following distance from surrounding vehicles.

Fig. 6: *Talk2BEV* in **free-form conversation** with a user. There is a car in front of the ego-vehicle (highlighted in red), which is reversing to park in a parking spot. *Talk2BEV* identifies the parking lights are on, and based on this visual information, and the spatial location of the car in front, *Talk2BEV* deems it unsafe to continue moving forward.

BEV	LVLM	2-Wheeler	Cars	Trucks	Construction
LSS	BLIP-2	0.56	0.60	0.67	0.67
	InstructBLIP-2	0.52	0.58	0.73	0.61
	MiniGPT-4	0.48	0.59	0.67	0.72
	Average	0.52	0.59	0.69	0.67
GT	BLIP-2	0.56	0.60	0.68	0.67
	InstructBLIP-2	0.56	0.58	0.74	0.67
	MiniGPT-4	0.56	0.66	0.72	0.72
	Average	0.56	0.61	0.71	0.68

TABLE IV: **Object Category-wise Evaluation:** Performance of *Talk2BEV* with Language Enhanced Map constructed with different LVLMs (BLIP-2, InstructBLIP-2, MiniGPT-4) and BEV variants (LSS and GT) on queries q_{mcq} for different vehicle categories.

terms of both IoU and Distance error, Hence, incorporating Spatial Operators enhances Talk2BEV's capability to tackle spatial reasoning challenges, providing the LLM with contextual depth and directing its attention to relevant components. **Object Category-wise Evaluation** From Table IV, it is evident that 2-wheeler vehicles, including bicycles and motorcycles, consistently showed lower performance compared to other categories. This is mainly due to their smaller BEV segmentation predictions, making it more difficult to accurately back-project when there are minor inconsistencies in the predicted positions. On the contrary, larger vehicles such as trucks and construction vehicles consistently outperformed cars in most cases. This can be attributed to their larger BEV segmentations, which enable more accurate back projections.

VI. Conclusion

In this work, we presented *Talk2BEV*, a language interface to BEV maps used in autonomous driving systems. By drawing upon recent advances in LLMs and LVLMs, *Talk2BEV* caters to a variety of AD tasks, including, but not limited to, visual and spatial reasoning, predicting unsafe traffic interactions, and plotting recourse. We also introduced Talk2BEV-Bench, a benchmark for evaluating subsequent work in LVLMs for AD applications. While we continue to integrate large pretrained models into AD stacks, we also emphasize the need for safety and alignment research before these models are deployed into safety-critical AD stacks.

References

[1] J. Li, D. Li, S. Savarese, and S. Hoi, "Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models," 2023. 1, 2, 5

[2] D. Zhu, J. Chen, X. Shen, X. Li, and M. Elhoseiny, “Minigpt-4: Enhancing vision-language understanding with advanced large language models,” 2023. 1, 2, 5

[3] H. Liu, C. Li, Q. Wu, and Y. J. Lee, “Visual instruction tuning,” 2023. 1, 2

[4] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, Y. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, A. Castro-Ros, M. Pellat, K. Robinson, D. Valter, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, “Scaling instruction-finetuned language models,” 2022. 1, 2

[5] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, and I. Stoica, “Judging llm-as-a-judge with mt-bench and chatbot arena,” 2023. 1, 2

[6] OpenAI. (2021) Chatgpt. Accessed: yyyy-mm-dd. [Online]. Available: https://www.openai.com/ 1, 2

[7] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M.-A. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, and G. Lample, “Llama: Open and efficient foundation language models,” 2023. 1, 2

[8] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, “Llama 2: Open foundation and fine-tuned chat models,” 2023. 1, 2

[9] OpenAI, “Gpt-4 technical report,” 2023. 1, 2, 4

[10] W. Dai, J. Li, D. Li, A. M. H. Tiong, J. Zhao, W. Wang, B. Li, P. Fung, and S. Hoi, “Instructblip: Towards general-purpose vision-language models with instruction tuning,” 2023. 1, 2, 5

[11] J. Philion and S. Fidler, “Lift, splat, shoot: Encoding images from arbitrary camera rigs by implicitly unprojecting to 3d,” 2020. 1, 2, 3

[12] Z. Li, W. Wang, H. Li, E. Xie, C. Sima, T. Lu, Q. Yu, and J. Dai, “Bevformer: Learning bird’s-eye-view representation from multi-camera images via spatiotemporal transformers,” 2022. 1, 2

[13] A. Hu, Z. Murez, N. Mohan, S. Dudas, J. Hawke, V. Badrinarayanan, R. Cipolla, and A. Kendall, “Fiery: Future instance prediction in bird’s-eye view from surround monocular cameras,” 2021. 1

[14] S. Hu, L. Chen, P. Wu, H. Li, J. Yan, and D. Tao, “St-p3: End-to-end vision-based autonomous driving via spatial-temporal feature learning,” 2022. 1, 2

[15] C. Fu, P. Chen, Y. Shen, Y. Qin, M. Zhang, X. Lin, Z. Qiu, W. Lin, J. Yang, X. Zheng, K. Li, X. Sun, and R. Ji, “Mme: A comprehensive evaluation benchmark for multimodal large language models,” 2023. 2

[16] Y. Liu, H. Duan, Y. Zhang, B. Li, S. Zhang, W. Zhao, Y. Yuan, J. Wang, C. He, Z. Liu, K. Chen, and D. Lin, “Mmbench: Is your multi-modal model an all-around player?” 2023. 2

[17] P. Xu, W. Shao, K. Zhang, P. Gao, S. Liu, M. Lei, F. Meng, S. Huang, Y. Qiao, and P. Luo, “Lvlm-ehub: A comprehensive evaluation benchmark for large vision-language models,” 2023. 2

[18] B. Li, R. Wang, G. Wang, Y. Ge, Y. Ge, and Y. Shan, “Seed-bench: Benchmarking multimodal llms with generative comprehension,” 2023. 2, 4

[19] P. Achlioptas, A. Abdelreheem, F. Xia, M. Elhoseiny, and L. J. Guibas, “Referit3d: Neural listeners for fine-grained 3d object identification in real-world scenes,” in *European Conference on Computer Vision*, 2020. [Online]. Available: https://api.semanticscholar.org/CorpusID:221378802 2

[20] P.-H. Huang, H.-H. Lee, H.-T. Chen, and T.-L. Liu, “Text-guided graph neural networks for referring 3d instance segmentation,” in *AAAI Conference on Artificial Intelligence*, 2021. [Online]. Available: https://api.semanticscholar.org/CorpusID:235306096 2

[21] M. Feng, Z. Li, Q. Li, L. Zhang, X. Zhang, G. Zhu, H. Zhang, Y. Wang, and A. Mian, “Free-form description guided 3d visual graph network for object grounding in point cloud,” 2021. 2

[22] D. Z. Chen, A. X. Chang, and M. Nießner, “Scanrefer: 3d object localization in rgb-d scans using natural language,” 2020. 2

[23] D. Z. Chen, A. Gholami, M. Nießner, and A. X. Chang, “Scan2cap: Context-aware dense captioning in rgb-d scans,” 2020. 2

[24] D. Azuma, T. Miyanishi, S. Kurita, and M. Kawanabe, “Scanqa: 3d question answering for spatial scene understanding,” 2022. 2

[25] S.-H. Chou, W.-L. Chao, W.-S. Lai, M. Sun, and M.-H. Yang, “Visual question answering on 360-degree images,” 2020. 2

[26] E. Wijmans, S. Datta, O. Maksymets, A. Das, G. Gkioxari, S. Lee, I. Essa, D. Parikh, and D. Batra, “Embodied question answering in photorealistic environments with point cloud perception,” 2019. 2

[27] X. Yan, Z. Yuan, Y. Du, Y. Liao, Y. Guo, Z. Li, and S. Cui, “Comprehensive visual question answering on point clouds through compositional scene manipulation,” *arXiv preprint arXiv:2112.11691*, 2021. 2

[28] Y. Hong, H. Zhen, P. Chen, S. Zheng, Y. Du, Z. Chen, and C. Gan, “3d-llm: Injecting the 3d world into large language models,” 2023. 2

[29] R. Xu, X. Wang, T. Wang, Y. Chen, J. Pang, and D. Lin, “Pointllm: Empowering large language models to understand point clouds,” 2023. 2

[30] S. N N, T. Maniar, J. Kalyanasundaram, V. Gandhi, B. Bhowmick, and M. Krishna, “Talk to the vehicle: Language conditioned autonomous navigation of self driving cars,” 11 2019, pp. 5284–5290. 2

[31] T. Deruyttere, S. Vandenhende, D. Grujicic, L. V. Gool, and M.-F. Moens, “Talk2car: Taking control of your self-driving car,” in *Conference on Empirical Methods in Natural Language Processing*, 2019. [Online]. Available: https://api.semanticscholar.org/CorpusID:202734592 2

[32] D. Wu, W. Han, T. Wang, Y. Liu, X. Zhang, and J. Shen, “Language prompt for autonomous driving,” 2023. 2

[33] A. B. Vasudevan, D. Dai, and L. V. Gool, “Object referring in videos with language and human gaze,” 2018. 2

[34] M. Cordts, M. Omran, S. Ramos, T. Rehfeld, M. Enzweiler, R. Benenson, U. Franke, S. Roth, and B. Schiele, “The cityscapes dataset for semantic urban scene understanding,” 2016. 2

[35] H. Caesar, V. Bankiti, A. H. Lang, S. Vora, V. E. Liong, Q. Xu, A. Krishnan, Y. Pan, G. Baldan, and O. Beijbom, “nuscenes: A multimodal dataset for autonomous driving,” 2020. 2

[36] D. Wu, W. Han, T. Wang, X. Dong, X. Zhang, and J. Shen, “Referring multi-object tracking,” 2023. 2

[37] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, “Roberta: A robustly optimized bert pretraining approach,” 2019. 2

[38] T. Qian, J. Chen, L. Zhuo, Y. Jiao, and Y.-G. Jiang, “Nuscenes-qa: A multi-modal visual question answering benchmark for autonomous driving scenario,” 2023. 2

[39] K. Mani, S. Daga, S. Garg, S. Shankar, K. Jatavallabhula, and M. K, “Monolayout: Amodal scene layout from a single image,” in *WACV*, 2020. 2

[40] K. Mani, S. Shankar, K. Jatavallabhula, and M. K, “Autolay: Benchmarking monocular layout estimation,” in *IROS*, 2020. 2

[41] X. Zhao, W. Ding, Y. An, Y. Du, T. Yu, M. Li, M. Tang, and J. Wang, “Fast segment anything,” 2023. 3

[42] J. Wu, J. Wang, Z. Yang, Z. Gan, Z. Liu, J. Yuan, and L. Wang, “Grit: A generative region-to-text transformer for object understanding,” 2022. 4

[43] Y. Du, C. Li, R. Guo, X. Yin, W. Liu, J. Zhou, Y. Bai, Z. Yu, Y. Yang, Q. Dang, and H. Wang, “Pp-ocr: A practical ultra lightweight ocr system,” 2020. 4

[44] K. M. Jatavallabhula, A. Kuwajerwala, Q. Gu, M. Omama, T. Chen, S. Li, G. Iyer, S. Saryazdi, N. Keetha, A. Tewari, J. B. Tenenbaum, C. M. de Melo, M. Krishna, L. Paull, F. Shkurti, and A. Torralba, “Conceptfusion: Open-set multimodal 3d mapping,” 2023. 3

[45] H. W. Chung, L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei, “Scaling instruction-finetuned language models,” 2022. [Online]. Available: https://arxiv.org/abs/2210.11416 3

[46] W.-L. Chiang, Z. Li, Z. Lin, Y. Sheng, Z. Wu, H. Zhang, L. Zheng, S. Zhuang, Y. Zhuang, J. E. Gonzalez, I. Stoica, and E. P. Xing, “Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality,” March 2023. [Online]. Available: https://lmsys.org/blog/2023-03-30-vicuna/ 3