From Scenes to Semantics: PersianClevr for Bilingual 3D Visual Reasoning

Kianoosh Vadaei*
Department of Computer Engineering
University of Isfahan
Isfahan, Iran
k.vadaei@mehr.ui.ac.ir

Melika Shirian*
Department of Computer Engineering
University of Isfahan
Isfahan, Iran
melika.shi@mehr.ui.ac.ir

Arshia Hemmat*
Department of Computer Science
University of Oxford
Oxford, UK
amirarshia.hemmat@kellogg.ox.ac.uk

Mohammad Hassan Heydari†
Department of Computer Engineering
University of Isfahan
Isfahan, Iran
m.heydari@mehr.ui.ac.ir

Ali Mamanpoosh†
Department of Computer Engineering
University of Isfahan
Isfahan, Iran
ali.mnp.uni@mehr.ui.ai.ir

Afsaneh Fatemi
Department of Computer Engineering
University of Isfahan
Isfahan, Iran
a_fatemi@eng.ui.ac.ir

Abstract

Vision–language models (VLMs) have made rapid progress on 2D visual reasoning, yet robust three-dimensional (3D) understanding and multilingual generalisation (particularly in Persian) remain underexplored. To address this gap, we introduce PersianClevr, a bilingual (English–Persian) benchmark targeting 3D scene understanding across five reasoning skills: attribute identification, counting, comparison, spatial relationships, and logical operations. PersianClevr is constructed by unifying CLEVR, Super-CLEVR, and ClevrTex; we synthesize missing question–answer pairs for ClevrTex with an instruction-tuned vision LLM and categorise items using an automated pipeline, then translate and balance the full set to yield parallel English–Persian splits. We outline evaluation protocols that test instructed VLMs in zero-shot and in-context learning settings, and include standard text metrics (BLEU, METEOR, ROUGE) for assessing translation quality alongside task accuracy. Together, these components provide a controlled, multilingual testbed for diagnosing compositional and spatial reasoning in 3D. We present baseline experiments and analyses to chart current strengths and failure modes, explicitly positioning PersianClevr as an evaluation resource for 3D-aware, multilingual VLMs rather than a new modeling technique.[3]

*Equal first authors.
†Equal second authors..
[3]dataset: https://huggingface.co/datasets/PrismaticLab/PersianClevr

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Workshop: NeurIPS 2025 VLM4RWD Workshop.

1 Introduction

Visual understanding in artificial intelligence (AI) has evolved rapidly, driven by advances in deep learning and multimodal architectures that integrate vision and language. As AI systems increasingly engage with complex visual tasks, vision-language models (VLMs) have emerged as pivotal frameworks, allowing machines to interpret and reason over both textual and visual inputs TG et al. [2024]. Despite significant progress in object recognition, spatial reasoning, and scene understanding, these models still struggle with fundamental challenges related to three-dimensional (3D) perception and real-world scene complexity Zhang et al. [2024]. The reliance on two-dimensional (2D) feature representations remains a bottleneck, as such representations fail to fully capture depth, occlusion, and intricate spatial relationships that define real-world environments Wang et al. [2024b]. While models have excelled in benchmark datasets for tasks such as visual question answering (VQA) and image captioning, their ability to generalize across diverse visual domains remains an open question Ghosh et al. [2024]. Furthermore, despite increasing interest in multilingual AI, the interaction between linguistic variation and visual reasoning remains insufficiently explored Han et al. [2024]. These gaps highlight the need for systematic evaluation frameworks that account for both scene complexity and linguistic diversity, and our work contributes such a framework in the form of a benchmark rather than a new model, ultimately enabling a more holistic understanding of how AI models perceive and process the visual worldKodali and Berleant [2022].

Figure 1: Overview figure illustrating example questions, categories, and the CLEVR family wheel.

Building on this foundation, recent efforts in multilingual vision-language models have showcased notable progress in extending robust visual understanding to a wider range of languages and cultural contexts. Yet, the ability to handle three-dimensional scenes—where intricate depth cues, occlusions, and spatial relationships play a pivotal role—remains relatively unexplored, particularly for Persian. This gap is most apparent in tasks like visual question answering (VQA), where models must integrate linguistic comprehension with the complexities of 3D visual reasoning, including counting objects, assessing occlusions, and inferring detailed spatial arrangements Delitzas et al. [2023], Parelli et al. [2023], Dwedari et al. [2023], Yang et al. [2024]. While proprietary or private models have demonstrated commendable performance on simpler Persian datasets, they have not been systematically tested on any comprehensive benchmark that targets 3D scene understanding. As a result, there is a pressing need for specialized evaluation frameworks that can rigorously assess how well different model architectures—and by extension, different linguistic capabilities—transfer to the nuanced challenges of 3D perception in Persian. Addressing this gap would not only advance the state of the art in multilingual AI but also foster a more inclusive global community equipped to tackle real-world visual tasks that transcend the limitations of traditional 2D benchmarks Fu et al. [2024], Li et al. [2023a], Man et al. [2024]. An overview of the four constituent sources and example bilingual QA pairs is shown in Fig. 1.

2 Background and Related Works

2.1 3D Scene Representation and Illusion Understanding

3D scene comprehension spans compositional reasoning, spatial inference, and multi-step logical composition. Benchmarks such as CLEVR3D Yan et al. [2023] and Ref-CLEVR Liu et al. [2019] examine object relations and attributes under controlled programs. Beyond compositional QA, the broader 3D literature covers semantic segmentation, detection, and scene reconstruction (e.g., Dense Multimodal Alignment Li et al. [2024]) and robotics-oriented reasoning about occlusion and affordances (e.g., Dream2Real Kapelyukh et al. [2024]). **IllusionBench** Hemmat et al. [2024] complements these by isolating *abstract shape recognition* under visual confounds, revealing that VLMs often rely on appearance shortcuts. **PersianClevr** extends this line toward *3D compositional reasoning* with controlled variation in viewpoint, occlusion, and texture, and adds a bilingual (EN/FA) axis to analyze language–vision interactions.

2.2 Persian-Language Contributions in Scene Understanding

Though most 3D scene understanding research has been conducted in English, recent efforts have begun to develop VLM resources for Persian. ParsVQA-Caps Mobasher et al. [2022] established a baseline for Persian VQA and captioning, highlighting linguistic challenges in multimodal settings. **MEENA** Ghahroodi et al. [2025] introduces *multimodal–multilingual educational exams* with n-level assessment, expanding coverage of Persian in evaluation and enabling controlled difficulty scaling across tasks. Beyond general Persian multimodal benchmarks, studies such as *Persian in a Court* Farsi et al. [2025] probe model disparities between Persian and high-resource languages. Yet, dedicated *3D* Persian benchmarks remain scarce, underscoring the need for datasets that explicitly test compositional, spatial, and logical reasoning under controlled 3D scenes—precisely the gap addressed by PersianClevr.

2.3 3D Question Answering from 2D Images

Recovering 3D spatial information from single images for QA has progressed from early 2D relation extraction to multi-view and 3D-aware approaches. 3D-Aware VQA targets part, pose, and occlusion queries using multimodal cues Wang et al. [2024a], while contrastive pretraining aligns 2D images, 3D point clouds, and text to improve scene grounding (Multi-CLIP Delitzas et al. [2023]; CLIP-Guided Pretraining for 3D QA Parelli et al. [2023]). On the data side, ScanQA introduces real-scene spatial QA expressly designed to bridge 2D–3D reasoning Azuma et al. [2022]. Together, these directions—paired with scaling vision–language models—push toward holistic 3D QA from predominantly 2D evidence Wang et al. [2024a], Delitzas et al. [2023], Parelli et al. [2023], Azuma et al. [2022].

2.4 Compositional and 3D Understanding Tasks

Compositional 3D reasoning spans object–attribute queries, relations, and spatial inference. CLEVR3D and Ref-CLEVR evaluate relational and reference understanding under controlled programs Yan et al. [2023], Liu et al. [2019]. Beyond QA, broader 3D understanding includes semantic segmentation, detection, and reconstruction, explored via dense multimodal alignment Li et al. [2024]. For robotics-oriented settings, reasoning under occlusion and learning object affordances are central (e.g., Dream2Real) Kapelyukh et al. [2024]. These benchmarks and tasks jointly chart the capabilities required for robust multimodal systems operating in real environments Yan et al. [2023], Liu et al. [2019], Li et al. [2024], Kapelyukh et al. [2024].

3 Dataset

As previously discussed, proposing the PersianClevr-Bench for vision-language models on multilingual 3D scene understanding comprises several key steps.

Dataset	Object Complexity	Texture Complexity	Scene Complexity
CLEVR Johnson et al. [2017]	Simple	Simple	Simple
CLEVR-Humans Johnson et al. [2017]	Simple	Simple	Simple
Super-CLEVR Li et al. [2023b]	Complex	Moderate	Complex
ClevrTex Karazija et al. [2021]	Simple	High	Moderate

Table 1: Complexity comparison across CLEVR, CLEVR-Humans (human-written questions over CLEVR images), Super-CLEVR, and ClevrTex.

3.1 Dataset Gathering and Preparation

We construct a multi-task English 3D scene-understanding benchmark by integrating three established datasets—CLEVR, Super-CLEVR, and ClevrTex Johnson et al. [2017], Li et al. [2023b], Karazija et al. [2021]. These datasets differ in visual complexity and object variability: CLEVR offers simple geometric objects with controlled attributes for evaluating compositional reasoning; Super-CLEVR introduces more complex 3D vehicle models that enable examination of domain shifts; and ClevrTex adds diverse, richly detailed textures that broaden the visual variability of the benchmark.

To ensure that the resulting benchmark functions as an evaluation-oriented resource, we categorize all samples based on their reasoning objectives and visual characteristics, then sample the final benchmark to maintain balanced coverage across reasoning types and dataset variants. This categorization supports a comprehensive set of 3D reasoning skills—including attribute identification, counting, comparison, spatial relations, and logical composition Shnitzer et al. [2023], Johnson et al. [2017]. Microsoft Phi-3.5 Abdin et al. [2024] serves as the base model for guiding this categorization process, and the combined visual diversity of the datasets provides a broad and controlled foundation for constructing the benchmark (see Appendix C).

3.1.1 Question Generation

Because ClevrTex Karazija et al. [2021] lacks question–answer annotations, we synthesized QA pairs for our sampled subset using the `o3` model, conditioning on each image and only *high-level* scene metadata (e.g., category such as `outd`) while deliberately omitting fine-grained attributes (size, position, geometry) to force reliance on visual evidence and better probe 3D understanding. We employed few-shot prompting with a small set of illustrative exemplars drawn from other sources in the benchmark to enforce a consistent reasoning format and encourage diverse, compositional question types. All QA were first produced in English and then translated into Persian to ensure linguistic parity, and the resulting answers serve as the ground-truth references used in our evaluations (see Section 4.2), yielding a coherent bilingual extension of ClevrTex aligned with the broader PersianClevr framework.

3.1.2 Dataset Translation

Thus far, the assembled dataset is entirely in English. To create a bilingual benchmark, we translated the full corpus into Persian and merged it with the original, yielding a multilingual 3D scene-understanding suite. Translation in this setting is non-trivial: questions encode compositional programs (attributes, relations, counting, logic), Persian exhibits morphological and word-order divergences from English, and several cues (e.g., left/right under changing viewpoints, material vs. texture terms, digit systems) must remain semantically and programmatically consistent with the images. To address this challenge, we employed *GPT-O3*, a reasoner model, to produce high-fidelity Persian translations aligned with the original QA semantics and answer formats. We used constrained JSON I/O with schema checks (one-word answers when required, fixed label inventories), numeric/orthographic normalization for Persian digits, and invariants on entity sets (color/shape/material/size) to maintain exact correspondence with the underlying scene programs. Automatic consistency guards (back-translation spot checks, contradiction tests, and answerability checks against the scene/program) were followed by targeted human post-editing on flagged items.

3.1.3 Model Answer Generation

We generated model answers in both Persian and English with language-specific strategies. For the Persian split, we used GPT-4o and Gemini-2 Flash under zero-shot and few-shot prompting, yielding four answer sets per image–question pair. In zero-shot, the prompt contained only the question and image with an explicit one-word answer constraint and no explanations. In few-shot, we prepended a small set of category-matched exemplars (each with an image, Persian question, and correct one-word answer), using disjoint images to avoid leakage, then queried the target under the same one-word constraint. For the English split, answers were produced only with Gemini-2 Flash in zero-shot; these results serve as a reference for cross-lingual comparison while the benchmark's primary focus remains Persian (see Section 4.2).

3.2 Dataset Statistics

The PERSIANCLEVR benchmark contains 40,000 question–answer pairs: 30,000 English examples generated using the methods in Section 3, and 10,000 Persian examples. Of these, 10,000—across both languages—include model-generated answers used specifically for evaluation. Our experiments focus on the 10,000-question Persian subset, which corresponds to 4,944 images and spans 49 question families. Each question is linked to a functional program averaging 10.96 steps (median 10, maximum 24), reflecting varied reasoning complexity. Question lengths remain concise, averaging 15.76 tokens in English and 14.05 in Persian.

The dataset covers a wide range of reasoning skills, often combining several within a single question. In the evaluated subset, comparison appears most frequently with 5,084 questions (72.6%), followed by counting in 4,826 (68.9%). Spatial reasoning occurs in 2,464 examples (35.2%), while attribute identification and logical operations are present in 1,101 (15.7%) and 526 (7.5%) questions, respectively. This distribution highlights the benchmark's emphasis on compositional, multi-step reasoning across complex 3D scenes.

3.3 Dataset Difficulty

We characterize difficulty along three axes—object/shape complexity, texture complexity, and scene complexity (clutter/occlusion)—as summarized in Table 1. In brief, CLEVR is low on all three (simple primitives, clean materials, sparse layouts), Super-CLEVR increases difficulty mainly through more complex objects (vehicles) and denser scenes with moderate texture variation, and ClevrTex keeps simple geometry but introduces high texture diversity with moderate scene clutter. We use this gradation when interpreting domain-wise results (shapes/vehicles/textures) and analyzing error patterns across languages and categories.

3.4 Experiment Overview

Models and language coverage. For the evaluation phase, all model-generated answers in both Persian and English were assessed using the `GPT-4o-mini` model. This configuration ensures a consistent and unbiased evaluation protocol across languages and models. Because reliable open-source VLMs with robust Persian (FA) instruction-following and answer-formatting capabilities are still limited, our primary focus remained on the Persian split. The evaluation therefore centers on comparing the model-generated Persian answers with their English counterparts to assess cross-lingual visual reasoning consistency.

Preprocessing. Prior to evaluation, all answers underwent a standardized preprocessing pipeline to ensure fairness and lexical uniformity across languages. For both Persian and English subsets, all numerical values were converted to their textual equivalents (e.g., "3" $\rightarrow$ "Persian word for three" and "3" $\rightarrow$ "three") to eliminate format-based discrepancies. Additionally, for binary and logical responses, a normalization procedure was applied to unify synonymous affirmative and negative expressions under a single canonical token. Two normalizations were applied for English and Persian, where positive expressions (*yes*, *correct*, *true*, *right*) were standardized as *yes*, and negative expressions (*no*, *incorrect*, *false*, *wrong*, *not true*, *not correct*) were standardized as *no*. This preprocessing ensured language-consistent evaluation and minimized semantic ambiguity in scoring.

Evaluation settings. We consider two evaluation settings: (i) *Zero-shot 3D understanding*, where the model receives only the scene image and a task instruction with no exemplars; and (ii) *In-context learning (ICL, Persian-only)*, where we prepend a small number of task-aligned exemplars ("shots") to the prompt so the model can infer the expected reasoning style and answer format. In ICL, exemplars are category-matched (same reasoning family as the query) and use disjoint images to prevent leakage; they are formatted to make the typically one-word answer explicit.

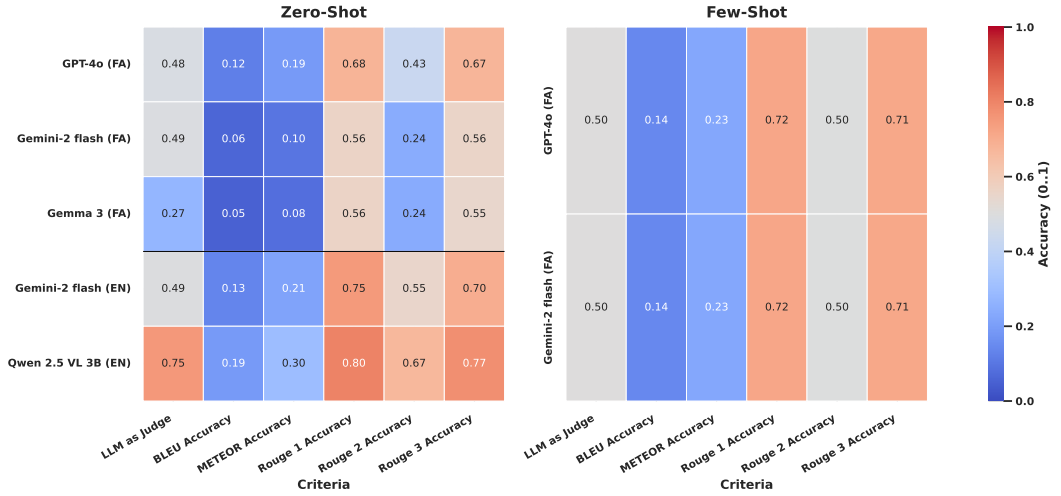

Figure 2: Comparison of model accuracies across evaluation metrics in zero-shot and few-shot settings.

4 Zero- and Few-Shot 3D Reasoning with Instructed VLMs

4.1 Experimental Setup

We evaluate instructed VLMs under two prompting tracks denoted by $\mathcal{T}_0$ (Zero-shot) and $\mathcal{T}_5$ (Few-shot with $K{=}5$ exemplars). Let $\mathcal{S} = \{\text{CLEVR}, \text{Super-CLEVR}, \text{ClevrTex}\}$ and $\mathcal{L} = \{\text{en}, \text{fa}\}$. Each source $S \in \mathcal{S}$ provides a set of image–question–answer triples in language ℓ given by

$$\mathcal{D}_{S,\ell} = \big\{(x_i, q_{i,\ell}, a_{i,\ell})\big\}_{i=1}^{n_{S,\ell}}. \tag{1}$$

The pooled zero-shot test set is

$$\mathcal{D}_{\text{ZS},\ell} = \bigcup_{S \in \mathcal{S}} \mathcal{D}_{S,\ell}. \tag{2}$$

An instructed VLM f_θ maps images and text prompts to answers. In $\mathcal{T}_0$ we apply a language-specific instruction template $\tau_\ell(\cdot)$ and predict as

$$\hat{a}_{i,\ell}^{(0)} = f_\theta\big(x_i, \tau_\ell(q_{i,\ell})\big). \tag{3}$$

In $\mathcal{T}_5$ we prepend a fixed, category-matched exemplar set $\mathcal{E}_{S,c,\ell}^{(5)} = \{(x_j^{(k)}, q_{j,\ell}^{(k)}, a_{j,\ell}^{(k)})\}_{k=1}^{5}$ and construct the prompt with $\Phi_\ell(\cdot,\cdot)$

$$\hat{a}_{i,\ell}^{(5)} = f_\theta\big(x_i, \Phi_\ell(\mathcal{E}_{S,c,\ell}^{(5)}, q_{i,\ell})\big). \tag{4}$$

Equations (1)–(4) define the datasets and predictors used throughout.

4.2 Evaluation

We report BLEU-2 Papineni et al. [2002], METEOR Banerjee and Lavie [2005], ROUGE-1/2/L Lin [2004], and an auxiliary LLM-as-a-Judge score following Liu et al. [2023], Zheng et al. [2023]. For the judge, item i in language ℓ receives $J(a_{i,\ell}, \hat{a}_{i,\ell}; q_{i,\ell}) \in \{0, \ldots, 100\}$ and we aggregate via

$$\overline{\text{Judge}}(\ell) = \frac{1}{N_\ell} \sum_{i=1}^{N_\ell} \frac{J(a_{i,\ell}, \hat{a}_{i,\ell}; q_{i,\ell})}{100}. \tag{5}$$

Table 2: Zero-shot metrics

Model	LLM as Judge	BLEU Acc.	METEOR Acc.	ROUGE-1 Acc.	ROUGE-2 Acc.	ROUGE-3 Acc.
GPT-4o (FA)	0.482	0.120	0.190	0.679	0.429	0.672
Gemini-2 Flash (FA)	0.490	0.061	0.096	0.562	0.241	0.556
Gemma 3 (FA)	0.272	0.051	0.080	0.559	0.238	0.545
Gemini-2 Flash (EN)	0.495	0.135	0.214	0.747	0.547	0.697
Qwen 2.5 VL 3B (EN)	0.751	0.189	0.297	0.802	0.668	0.774

Table 3: Few-shot metrics

Model	LLM as Judge	BLEU Acc.	METEOR Acc.	ROUGE-1 Acc.	ROUGE-2 Acc.	ROUGE-3 Acc.
GPT-4o (FA)	0.504	0.143	0.226	0.719	0.499	0.714
Gemini-2 Flash (FA)	0.504	0.143	0.226	0.719	0.499	0.714

We reference Eq. (5) when reporting judge means per language. All metrics use shared normalization (case, whitespace, Persian/Arabic digits).

4.3 Zero-shot Results and Model Inventory ($\mathcal{T}_0$)

Persian (fa). We evaluate **GPT-4o**, **Gemini-2 Flash**, and open-source **Gemma 3-4B**. Private VLMs are prioritized for Persian due to the scarcity of strong Persian-capable open models.

English (en). We run zero-shot on **Gemma 3-4B**, **Gemini-2 Flash**, and **Qwen-2.5 VL-3B**.

Findings. Table 2 reports corpus-level means. Persian private VLMs outperform the small open baseline across BLEU, METEOR, ROUGE, and the judge, while English zero-shot generally scores higher on overlap metrics. These trends, together with Eq. (3), motivate adding contextual guidance for Persian.

4.4 Few-shot Results ($\mathcal{T}_5$; $K{=}5$)

Scope. Few-shot is applied to **Persian** only with **GPT-4o** and **Gemini-2 Flash**. We use five fixed, relevant shots per dataset–category, with disjoint images and one-word answers.

Rationale and construction. Given the zero-shot gaps in Table 2, we introduce $\mathcal{T}_5$ to provide light format guidance without any model fine-tuning. Prediction follows Eq. (4) with a fixed exemplar set per bucket. As shown in Table 3, both GPT-4o and Gemini-2 Flash improve over Persian zero-shot across reported metrics, which we also summarize via the judge in Eq. (5).

5 Findings and Analysis on VLMs' 3D Visual Understanding

Evaluations on PersianCLEVR reveal that instructed VLMs demonstrate promising multimodal understanding yet continue to struggle when tasks require full three-dimensional reasoning or multi-step logic. As shown in Table 2, models achieve moderate performance under the zero-shot condition, with BLEU and METEOR scores generally ranging in the lower to middle bands, and an average semantic evaluation (LLM-as-Judge) below 0.50. When few-shot examples are provided (Table 3), all models show modest but consistent improvement, BLEU and METEOR scores increase slightly, while the LLM-as-Judge metric rises above 0.50. These gains confirm that exposure to examples helps models generate more structured responses, though not necessarily deeper reasoning. Across both languages, a noticeable gap remains between lexical and semantic metrics, showing that while outputs often appear fluent, they do not always reflect accurate reasoning.

5.1 Attributes

When objects are clearly visible and distinct, models identify properties such as shape, color, material, and size with reliable accuracy. Performance trends remain steady across both evaluation settings in Tables 2 and 3, suggesting that attribute recognition is one of the stronger and more stable aspects of multimodal understanding. However, accuracy declines when an attribute must be inferred through relational phrasing, e.g., "the object next to…" or "behind the cube." This indicates that although visual features are recognized effectively, connecting them with spatial context is still a source of error. The small difference between lexical and semantic scores supports the observation that errors are mainly due to incomplete spatial grounding rather than linguistic variability.

5.2 Counting and Comparison

Counting tasks perform well for small quantities and simple visual scenes but degrade as complexity increases. Comparative questions (greater/less/equal) emphasize this limitation: even small errors in counting can reverse the final answer. As indicated by the improvements between Table 2 and Table 3, BLEU and METEOR scores rise slightly in the few-shot setting, showing that examples assist models in recognizing basic numeric patterns. Nevertheless, the improvements remain marginal, suggesting that models still find it difficult to maintain object tracking and accurate set representations when multiple visual and spatial references overlap.

5.3 Spatial Relations

Spatial reasoning continues to be the most challenging component of 3D understanding. Questions involving depth, occlusion, or multiple reference points result in the largest drop in accuracy. Although Table 3 shows small lexical gains compared with Table 2, these changes likely reflect better wording rather than genuine reasoning progress. The results suggest that models rely heavily on planar, two-dimensional cues, interpreting relations such as "in front of" or "behind" based on visible layout rather than true spatial depth. This leads to partially correct but semantically incomplete answers, confirming that current architectures still lack explicit volumetric reasoning mechanisms.

5.4 Logical Composition

Questions combining several reasoning steps, like conjunctions, disjunctions, or negations, expose limited compositional structure within current models. While responses often include correct elements, they rarely satisfy all components of a question simultaneously. This trend is visible in both zero-shot and few-shot metrics (Tables 2 and 3), where lexical scores show modest improvements but semantic consistency lags behind. These findings highlight the difficulty models face in connecting perception with higher-order symbolic reasoning, showing that lexical overlap alone is not a reliable measure of logical accuracy.

5.5 Cross-Lingual Behavior

Matched zero-shot evaluations favor English on BLEU/METEOR, while the semantic judge differs only slightly (from just under to slightly above 0.50; Table 2). This suggests the disparity is driven more by script, tokenization, and morphology than core reasoning. Overall, lexical metrics are sensitive to language form, whereas semantic scoring better reflects multilingual reasoning.

5.6 Effect of In-Context Learning

Few-shot prompting yields modest but consistent improvements, with BLEU, METEOR, and LLM-as-Judge scores rising slightly (Tables 2 and 3), especially for attribute identification and counting. These gains mainly reflect better alignment with task phrasing rather than true reasoning advances. Spatial and logical questions remain difficult due to their reliance on multi-relation integration and depth cues, showing that current instructed VLMs still act largely as two-dimensional describers whose coherence breaks under compositional or geometric demands. Meaningful progress requires explicit 3D signals such as depth or multi-view supervision and evaluation metrics focused on reasoning accuracy, and future benchmarks should incorporate lexical and semantic measures along with controlled spatial variation to better assess genuine multi-modal reasoning.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024.

Daichi Azuma, Taiki Miyanishi, Shuhei Kurita, and Motoaki Kawanabe. Scanqa: 3d question answering for spatial scene understanding. In *proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 19129–19139, 2022.

Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pages 65–72, 2005.

Alexandros Delitzas, Maria Parelli, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Multi-clip: Contrastive vision-language pre-training for question answering tasks in 3d scenes. *arXiv preprint arXiv:2306.02329*, 2023.

Mohammed Munzer Dwedari, Matthias Niessner, and Dave Zhenyu Chen. Generating context-aware natural answers for questions in 3d scenes. *arXiv preprint arXiv:2310.19516*, 2023.

Farhan Farsi, Shahriar Shariati Motlagh, Shayan Bali, Sadra Sabouri, and Saeedeh Momtazi. Persian in a court: Benchmarking vlms in persian multi-modal tasks. In *Proceedings of the First Workshop of Evaluation of Multi-Modal Generation*, pages 52–56, 2025.

Rao Fu, Jingyu Liu, Xilun Chen, Yixin Nie, and Wenhan Xiong. Scene-llm: Extending language model for 3d visual understanding and reasoning. *arXiv preprint arXiv:2403.11401*, 2024.

Omid Ghahroodi, Arshia Hemmat, Marzia Nouri, Seyed Mohammad Hadi Hosseini, Doratossadat Dastgheib, Mohammad Vali Sanian, Alireza Sahebi, Reihaneh Zohrabi, Mohammad Hossein Rohban, Ehsaneddin Asgari, et al. Meena (persianmmmu): Multimodal-multilingual educational exams for n-level assessment. *arXiv preprint arXiv:2508.17290*, 2025.

Akash Ghosh, Arkadeep Acharya, Sriparna Saha, Vinija Jain, and Aman Chadha. Exploring the frontier of vision-language models: A survey of current methodologies and future directions. *arXiv preprint arXiv:2404.07214*, 2024.

Soyeon Caren Han, Feiqi Cao, Josiah Poon, and Roberto Navigli. Multimodal large language models and tunings: Vision, language, sensors, audio, and beyond. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 11294–11295, 2024.

Arshia Hemmat, Adam Davies, Tom Lamb, Jianhao Yuan, Philip Torr, Ashkan Khakzar, and Francesco Pinto. Hidden in plain sight: evaluating abstract shape recognition in vision-language models. *Advances in Neural Information Processing Systems*, 37:88527–88556, 2024.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 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, 2017.

Ivan Kapelyukh, Yifei Ren, Ignacio Alzugaray, and Edward Johns. Dream2real: Zero-shot 3d object rearrangement with vision-language models. In *First Workshop on Vision-Language Models for Navigation and Manipulation at ICRA 2024*, 2024.

Laurynas Karazija, Iro Laina, and Christian Rupprecht. Clevrtex: A texture-rich benchmark for unsupervised multi-object segmentation. *arXiv preprint arXiv:2111.10265*, 2021.

Venkat Kodali and Daniel Berleant. Recent, rapid advancement in visual question answering: a review. In *2022 IEEE International Conference on Electro Information Technology (eIT)*, pages 139–146. IEEE, 2022.

Lin Li, Haohan Zhang, and Zeqin Fang. An empirical study of multilingual scene-text visual question answering. In *Proceedings of the 2nd Workshop on User-centric Narrative Summarization of Long Videos*, pages 3–8, 2023a.

Ruihuang Li, Zhengqiang Zhang, Chenhang He, Zhiyuan Ma, Vishal M Patel, and Lei Zhang. Dense multimodal alignment for open-vocabulary 3d scene understanding. In *European Conference on Computer Vision*, pages 416–434. Springer, 2024.

Zhuowan Li, Xingrui Wang, Elias Stengel-Eskin, Adam Kortylewski, Wufei Ma, Benjamin Van Durme, and Alan L Yuille. Super-clevr: A virtual benchmark to diagnose domain robustness in visual reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14963–14973, 2023b.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

Runtao Liu, Chenxi Liu, Yutong Bai, and Alan L Yuille. Clevr-ref+: Diagnosing visual reasoning with referring expressions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4185–4194, 2019.

Yang Liu, Lianhui Xu, Wanyu Xiao, and et al. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

Yunze Man, Liang-Yan Gui, and Yu-Xiong Wang. Situational awareness matters in 3d vision language reasoning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13678–13688, 2024.

Shaghayegh Mobasher, Ghazal Zamaninejad, Maryam Hashemi, Melika Nobakhtian, and Sauleh Eetemadi. Parsvqa-caps: A benchmark for visual question answering and image captioning in persian. *people*, 101:404, 2022.

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

K Papinesi. Bleu: A method for automatic evaluation of machine translation. In *Proc. 40th Actual Meeting of the Association for Computational Linguistics (ACL), 2002*, pages 311–318, 2002.

Maria Parelli, Alexandros Delitzas, Nikolas Hars, Georgios Vlassis, Sotirios Anagnostidis, Gregor Bachmann, and Thomas Hofmann. Clip-guided vision-language pre-training for question answering in 3d scenes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5607–5612, 2023.

Tal Shnitzer, Anthony Ou, M'irian Silva, Kate Soule, Yuekai Sun, Justin Solomon, Neil Thompson, and M. Yurochkin. Large language model routing with benchmark datasets. *ArXiv*, abs/2309.15789, 2023. doi: 10.48550/arXiv.2309.15789.

Adithya TG, Adithya SK, Abhinav R Bharadwaj, Abhiram HA, and Surabhi Narayan. Enhancing vision models for text-heavy content understanding and interaction, 2024. URL `https://arxiv.org/abs/2405.20906`.

Xingrui Wang, Wufei Ma, Zhuowan Li, Adam Kortylewski, and Alan L Yuille. 3d-aware visual question answering about parts, poses and occlusions. *Advances in Neural Information Processing Systems*, 36, 2024a.

Zhecan Wang, Garrett Bingham, Adams Wei Yu, Quoc V Le, Thang Luong, and Golnaz Ghiasi. Haloquest: A visual hallucination dataset for advancing multimodal reasoning. In *European Conference on Computer Vision*, pages 288–304. Springer, 2024b.

Xu Yan, Zhihao Yuan, Yuhao Du, Yinghong Liao, Yao Guo, Shuguang Cui, and Zhen Li. Comprehensive visual question answering on point clouds through compositional scene manipulation. *IEEE Transactions on Visualization and Computer Graphics*, 2023.

Dejie Yang, Zhu Xu, Wentao Mo, Qingchao Chen, Siyuan Huang, and Yang Liu. 3d vision and language pretraining with large-scale synthetic data. *arXiv preprint arXiv:2407.06084*, 2024.

Huaxiang Zhang, Yaojia Mu, Guo-Niu Zhu, and Zhongxue Gan. Insightsee: Advancing multi-agent vision-language models for enhanced visual understanding. *arXiv preprint arXiv:2405.20795*, 2024.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, and et al. Mt-bench: Judging llm-as-a-judge with multi-turn, multi-facet evaluation. *arXiv preprint arXiv:2306.05685*, 2023.

A Dataset Documentation and Additional Information

A.1 Dataset Categorization

PersianClevr is a multi-modal bilingual 3D scene understanding benchmark in 5 different categories: attribute identification, counting, comparison, spatial relationships, and logical operations:

Attribute Identification Questions in this category test the ability to isolate and recognize properties of individual objects. This assesses perceptual abilities related to object recognition (identifying cubes, spheres, cylinders), attribute extraction (determining color, size, material, and shape), and focused attention to the relevant object while ignoring other parts of the image.

Counting Moving beyond simple object recognition, counting questions require object localization (identifying the positions of multiple objects), filtering or selection (applying criteria based on attributes to select a subset of objects), and quantity reasoning (determining the number of selected objects, demonstrating numerical understanding).

Comparison Comparison questions necessitate attribute identification and/or counting, but also working memory (temporarily storing information), relational reasoning (evaluating relationships between attributes or counts), and value retrieval (recalling and comparing information from memory). This includes determining if values or quantities are equal or unequal.

Spatial Relationships This category tests understanding of geometric relationships, requiring spatial understanding (interpreting spatial prepositions), relative positioning (determining the position of one object relative to another), and spatial reasoning (inferring spatial relationships, which may not be immediately obvious). The 3D relationships are dependent on the camera's viewpoint.

Logical Operations The most complex category, logical operations demands decomposition (breaking a question into subtasks), compositional reasoning (combining the results according to logical operators like AND or OR), abstract reasoning (reasoning over object attributes and relationships) and multi-step inference (performing sequences of reasoning steps).

After categorizing the dataset into the defined categories, we balanced the number of samples in each category to ensure a fair and comprehensive benchmark across all objectives.

A.2 ClevrTex Question Generation

ClevrTex introduces more complex textures to the objects in each scene, but it lacks corresponding questions for these scenes, rendering it fully unsupervised. To address this limitation, we incorporated a synthetic data generation step to create a supervised multi-modal dataset based on ClevrTex. This enriched dataset was then integrated into our proposed PersianClevr.

A.3 Licenses and Terms of Use

We respect and inherit the licenses of all upstream sources used to build *PersianClevr*. Below we list the license for each dataset we directly use to compose our benchmark splits.

Notes. (1) We release only derived QA pairs and Persian translations; original images remain governed by their upstream licenses. (2) Super-CLEVR's repository contains a `LICENSE` file that governs code and release assets; consult the repository for the most up-to-date terms. (3) For completeness, external datasets cited but not used to build our splits include *ScanQA* (CC BY-NC-SA 3.0) and *MEENA (PersianMMMU)* (CC BY-ND), which we reference in related work and comparisons.

Dataset	License / Terms
CLEVR	Creative Commons Attribution 4.0 (CC BY 4.0)[4]
Super-CLEVR	See project `LICENSE` for current terms (repository release)[5]
ClevrTex	Creative Commons Attribution 4.0 (CC BY 4.0)[6]

Table 4: Licenses for upstream datasets used to construct *PersianClevr*. We redistribute only our QA/translation annotations and do not mirror upstream imagery or meshes; users must comply with original licenses.

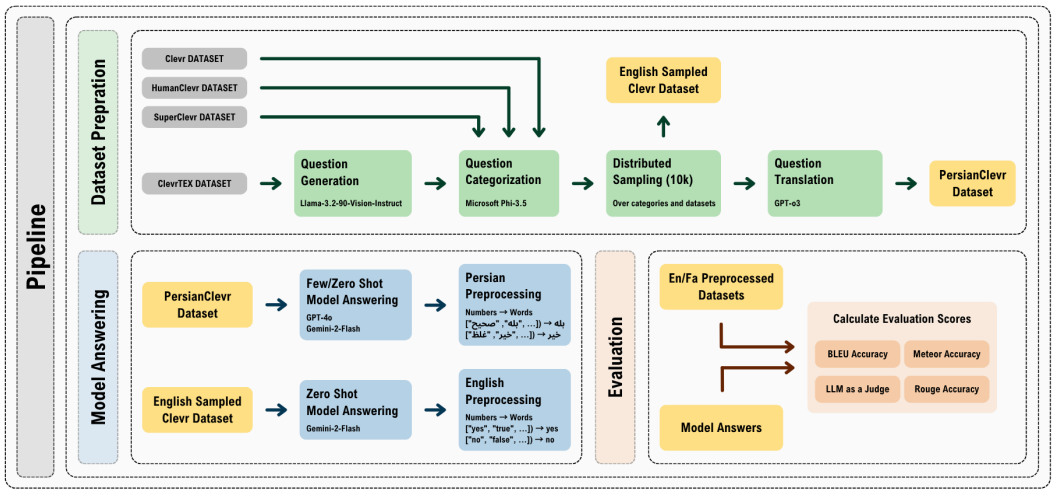

Figure 3: Overview of the dataset generation, model answering, and evaluation pipeline, showing how multiple PersianClevr dataset is processed, translated, and used to evaluate English and Persian visual question-answering models with automated and LLM-based scoring.

B Evaluations

B.1 Traditional Evaluation Metrics

BLEU BLEU (Bilingual Evaluation Understudy) [Papinesi, 2002] is a precision-based metric that evaluates translation quality by measuring the overlap of n-grams between the candidate translation and reference translations. The BLEU score is computed using the equation6.

$$\text{BLEU} = BP \times \exp\left(\sum_{n=1}^{N} w_n \log p_n\right) \tag{6}$$

where BP is the brevity penalty, p_n is the precision of n-grams, and w_n are weights assigned to different n-gram orders.

METEOR METEOR (Metric for Evaluation of Translation with Explicit ORdering) [Banerjee and Lavie, 2005] is a recall-oriented metric that considers synonym matching and stemming, making it more robust than BLEU in some cases. The METEOR score is computed using the equation 7

$$\text{METEOR} = F_{mean} \times (1 - P_{penalty}) \tag{7}$$

where F_{mean} is the harmonic mean of precision and recall, and $P_{penalty}$ is a fragmentation penalty to account for word order differences.

ROUGE ROUGE (Recall-Oriented Understudy for Gisting Evaluation) [Lin, 2004] is a recall-based metric commonly used for summarization but also applicable to translation. The most frequently used variant, ROUGE-N, is defined as equation 8.

$$\text{ROUGE-N} = \frac{\sum_{s \in \text{Ref}} \sum_{n-gram \in s} \text{match}(n-gram)}{\sum_{s \in \text{Ref}} \sum_{n-gram \in s} \text{count}(n-gram)} \tag{8}$$

Dataset	Modality	Language(s)	Task	Size (approx)	Open	Numeric UI	Diff./Trap?
PersianClevr	Img (Synth)	English, Persian	3D visual reasoning (bilingual VQA)	∼12k images, ∼40k QA pairs	Planned	Yes	Yes
CLEVR	Img (Synth)	English	3D VQA (compositional reasoning)	100k images, ∼1M QA pairs	Yes	Yes	Yes
Super-CLEVR	Img (Synth)	English	3D VQA (domain robustness)	30k images, ∼600k QA pairs	Yes	Yes	Yes
ClevrTex	Img (Synth)	–	Unsupervised multi-object segmentation	60k images (50k train + 10k test)	Yes	N/A	N/A
ParsVQA-Caps	Img (Real)	Persian (+English)	Multilingual VQA & captioning	Few thousand images; human and template QA	Yes	Yes	No
ScanQA	3D (RGB-D)	English	3D spatial QA (real scenes)	800 scenes; >41k Q–A pairs	Yes	Yes	No
Multi-CLIP	2D+3D	English	Pre-training (2D–3D contrastive VL)	Uses existing 3D scene data	N/A	N/A	N/A
CLIP-Guided 3D QA Pretraining	2D+3D	English	Pre-training (VL for 3D QA)	Uses existing 2D/3D data	N/A	N/A	N/A
CLEVR3D	3D (Point clouds)	English	3D VQA (compositional)	8,771 3D scenes; ∼171k Qs	Yes	Yes	Yes
CLEVR-Ref+	Img (Synth)	English	Referring expression comprehension	100k+ synthetic images	Yes	No	Yes
Super-CLEVR-3D	Img (Synth)	English	3D-aware VQA (parts, poses)	30k images; new part/pose Qs	Yes	Yes	Yes
MEENA (PersianMMMU)	Img (Real + Diagram)	Persian, English	Multilingual VQA (scientific, reasoning, educational)	∼10k Qs (7.5k Persian, 3k English); image-based MCQs	Yes	Yes	Yes

Table 5: Comparison of multimodal datasets for visual reasoning and VQA tasks (without notes column).

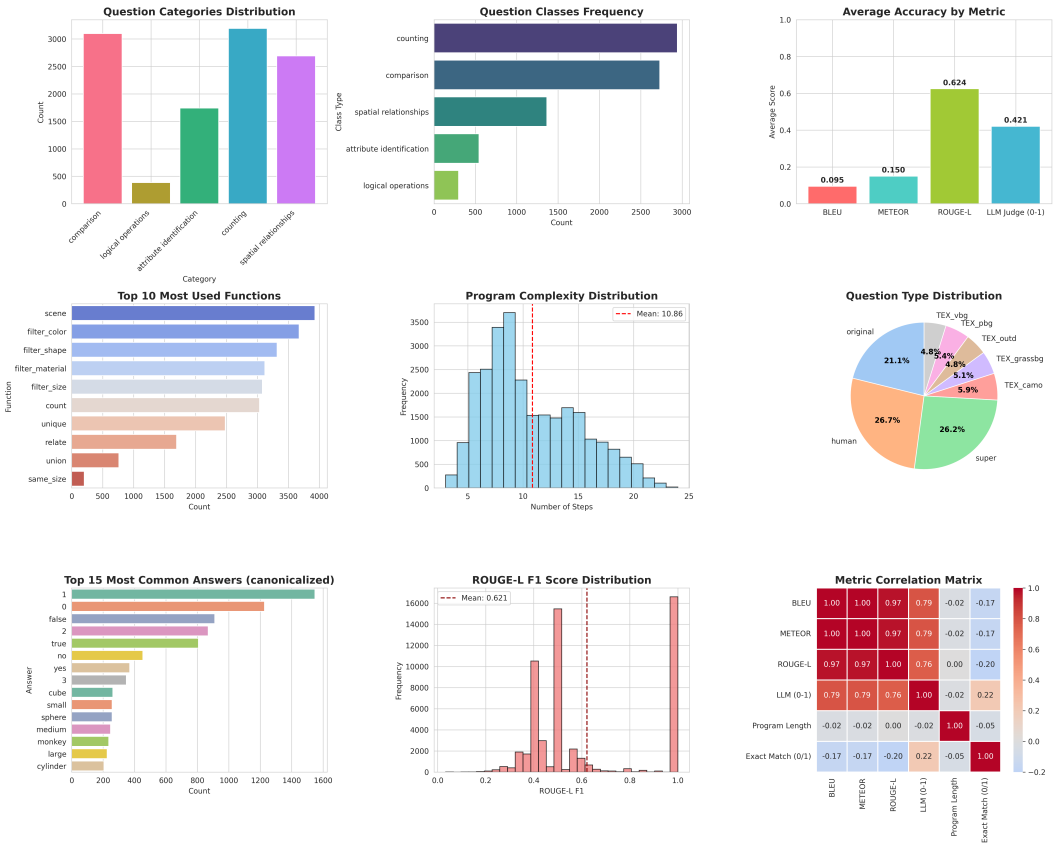

Figure 4: Summary of dataset characteristics and performance statistics for the Persian subset.

where the numerator counts the number of overlapping n-grams between candidate and reference translations, and the denominator is the total count of n-grams in the reference.

C Prompts

In this appendix section, we present all the prompts provided to the LLMs for each step in the dataset creation process.

Figure 5: Additional analysis of accuracy trends and distributional properties for the Persian subset.

C.1 Categorization Prompt

Categorization Prompt

You are given a query and must decide which of the following categories it falls into: - attribute identification - counting - comparison - spatial relationships - logical operations
A single query can have multiple categories. Output only the categories separated by commas, and do not include any extra text or explanations.
Here are some examples:
Q: Are there an equal number of large things and metal spheres?
A: comparison, counting, attribute identification
Q: How many objects are either small cylinders or red things?
A: logical operations, counting, attribute identification
Q: There is a sphere with the same size as the metal cube; is it made of the same material as the small red sphere?
A: attribute identification, comparison
Q: What size is the cylinder that is left of the brown metal thing that is left of the big sphere?
A: spatial relationships
Now your query is:
Q: {**QUESTION**}?
A:

Figure 6: Summary of dataset characteristics and performance statistics for the English subset.

C.2 Question Generation for ClevrTex

Question Generation for ClevrTex

You are a helpful assistant specializing in 3D scene understanding.
You will receive an image and must generate exactly five question-answer pairs in valid JSON format with the following ten keys:
question-1, answer-1, question-2, answer-2, question-3, answer-3, question-4, answer-4, question-5, and answer-5.
Each question must be complex by combining at least two of these 3D scene categories: attribute identification (e.g., color, shape, material), counting, comparison (e.g., bigger, smaller, taller), spatial relationships (e.g., left, right, behind), and logical operations (e.g., not, and, or).
At least three of the five questions must be WH-questions (e.g., "What," "Which," "Where," "How"), ensuring more WH-questions than yes/no questions. Every question must be answerable with exactly one word (e.g., "yes," "no," "red," "cube").
The final response must be strictly the JSON object with the specified ten keys and no additional commentary or formatting.

Analyze the following image and create five HARD question-answer pairs as specified.
Image : {**IMAGE**}

You are a helpful assistant specializing in 3D scene understanding. You will receive an image and must generate exactly five question-answer pairs in valid JSON format with the following ten keys: question-1, answer-1, question-2, answer-2, question-3, answer-3, question-4, answer-4, question-5, and answer-5.

Figure 7: Additional analysis of accuracy trends and distributional properties for the English subset.

Each question must be complex by combining at least two of these 3D scene categories: attribute identification (e.g., color, shape, material), counting, comparison (e.g., bigger, smaller, taller), spatial relationships (e.g., left, right, behind), and logical operations (e.g., not, and, or).

C.3 Dataset Translation

Dataset Translation

Translate the following pairs of English questions and answers into Persian accurately. Follow these steps to ensure precision:
1. Understand the Meaning: Read the English question and answer carefully to fully grasp their meaning. Pay attention to context, tone, and any nuances.
2. Identify Key Terms: Recognize important words, technical terms, or idiomatic expressions that may require special handling in Persian.
3. Preserve the Meaning: Ensure the translation conveys the exact same meaning without introducing any ambiguity or unintended changes. Do not perform word-for-word translation if it alters the meaning.
4. Maintain Fluency and Naturalness: The Persian translation should sound natural, fluent, and grammatically correct. Adapt sentence structures as needed to fit Persian language norms.
5. Verify Accuracy: After translating, compare the Persian version with the original English text to confirm that no meaning is lost or altered.
6. Output Format: Each translated question-answer pair must be in a VALID JSON format, with these keys: en-qa, en-ans, fa-qa, fa-ans
Example:
English Question: What number of other objects are there of the same size as the purple cube?
English Answer: 1
Persian Translation:
Now, translate the following pair of question and answer:
English Question: { **QUESTION** } English Answer: { **ANSWER** }

D Computational Resources and Access

No heavy-cluster access. We did not have access to institutional HPC clusters. All experiments were executed using commodity GPU notebooks and hosted inference APIs.

Kaggle GPUs (open-source baselines). Open-source VLM baselines (e.g., Llama-Vision, InternVL-4B) were run on *Kaggle* notebooks using NVIDIA **T4** and **P100** accelerators. Given VRAM limits, inference used batch size 1, mixed precision when available, and deterministic decoding. Where needed, images were resized once and cached; prompts and outputs were written to JSONL logs for exact replay.

Hosted inference (closed + open models). For hosted runs we used (i) HuggingFace Inference endpoints for open-source checkpoints and (ii) the AVALAI API provider for access to proprietary models (e.g., GPT-4o, Gemini-2 Flash). All API calls used fixed instruction templates, temperature $= 0$, and identical decoding limits per language to ensure comparability. We log request payloads (prompts, seeds, model IDs) and responses (raw text plus parsed answers) with timestamps to enable audit and re-evaluation.

Reproducibility controls. (i) Single, versioned prompt templates per language; (ii) deterministic decoding; (iii) seeding of any stochastic pre/post-processing; (iv) per-item caching of model I/O; (v) exact model identifiers and endpoint URIs recorded alongside results. These controls allow third parties to reproduce our zero-shot evaluations without access to large compute.

Why zero-shot on English (and open-source) only. Besides avoiding prompt-engineering confounds, zero-shot evaluation keeps compute and API usage bounded under our resource constraints, while still revealing intrinsic model capability. Persian ICL experiments are limited to API models to ensure stable formatting and avoid mixing language adaptation with visual reasoning under low-resource hardware.

E Limitation

While PersianClevr marks a meaningful advancement in bilingual 3D visual reasoning benchmarks, particularly for low-resource languages such as Persian, it is not without limitations in its dataset construction and overall scope. The benchmark's dependence on synthetic 3D scenes sourced from CLEVR, Super-CLEVR, and ClevrTex inherently constrains its applicability to real-world settings, as these scenes omit the inherent complexities of natural environments, including variations in lighting, sensor noise, and unstructured backgrounds, which could lead to overly optimistic assessments of model performance in applied domains. Furthermore, with a relatively modest scale of approximately 5k scenes and 10k question-answer pairs, the dataset offers limited coverage of edge cases and may not support comprehensive fine-tuning efforts. Its exclusive emphasis on five core reasoning categories (attribute identification, counting, comparison, spatial relationships, and logical operations) also precludes evaluation of more expansive 3D capabilities, such as semantic segmentation or affordance prediction, while the object repertoire remains confined to geometric primitives and vehicular models, thereby excluding more diverse entities like household objects or natural elements. Additionally, the project's scope was constrained by our available infrastructure and API limits, which restricted the breadth of model evaluations and the generation of larger-scale annotations.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes] , [No] , or [NA] .
- [NA] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for NA).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will be asked to also include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While "[Yes] " is generally preferable to "[No] ", it is perfectly acceptable to answer "[No] " provided a proper justification is given (e.g., "error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering "[No] " or "[NA] " is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

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

   Answer: [Yes]

   Justification: The abstract and introduction clearly state the paper's main contributions—introducing PersianClevr as a bilingual benchmark for 3D visual reasoning, describing its construction pipeline, and defining five reasoning categories.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

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

   Answer: [Yes]

Justification: The Appendix (Limitation section) outlines key limitations, including reliance on synthetic scenes, limited dataset scale, and restricted reasoning and object diversity.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

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

   Answer: [NA] .

   Justification: The paper contains no theoretical results or proofs; Sections 3–6 focus solely on empirical methods and evaluations, instead, it focuses on dataset construction, evaluation protocols, and empirical findings.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

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

   Answer: [Yes]

Justification: Sections 3.1 and 3.3, along with the Appendix, include detailed descriptions of dataset construction, evaluation setup, and all prompt templates, providing enough information for others to reproduce the benchmark and main experimental results.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Section 3.1 and the Appendix describe that the PersianClevr dataset and scripts will be released with instructions to reproduce the benchmark and experiments.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.

- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: Sections 3.3 and 4.1–4.5 clearly specify dataset splits, evaluation settings, prompting strategies, and model usage, providing sufficient detail to understand the experimental setup.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

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

   Answer: [No]

   Justification: This work is primarily a dataset/benchmark release. The tables include single-pass, deterministic baseline runs (fixed prompts, temperature=0, fixed decoding) provided only as reference points rather than hypothesis tests.

8. **Experiments compute resources**

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

   Answer: [Yes]

   Justification: Appendix A.8 specifies GPU type, memory, cloud provider, and total compute time used, providing sufficient reproducibility details.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: The work follows the NeurIPS Code of Ethics, using only synthetic and publicly available data with no human subjects, privacy risks, or potential for harmful use.

   Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [NA] .

    Justification: The work introduces a synthetic benchmark, which has no direct societal impact beyond methodological research.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

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

    Answer: [NA] .

    Justification: The paper poses no misuse risks, as it releases only synthetic benchmark data without real images, human content, or pretrained models requiring safeguards.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

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

Answer: [Yes]

Justification: The paper cites all major datasets used (CLEVR, Super-CLEVR, and ClevrTex) and provides proper credit. In Appendix (Dataset section), the authors explicitly mention that the datasets are used under the CC-BY 4.0 license, satisfying licensing and attribution requirements.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

Answer: [Yes]

Justification: Section 3.1 (Dataset Construction) and the Appendix document the new dataset, detailing its composition, generation process, and intended use.

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

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

Answer: [NA] .

Justification: The paper does not involve any crowdsourcing or research with human subjects; all data are synthetically generated and automatically processed.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.

- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

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

    Answer: [NA] .

    Justification: The research does not involve human subjects or participant studies, so no IRB or equivalent ethical review was required.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: Sections 3.1 describe the use of LLMs for generating question–answer pairs, categorizing items, and translating text as core components of the dataset construction process.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.