

## 849 A Model Prompts

850 For models trained for multi-image input, text prompt is:

851 Which objects are present in both images? Select all choices that are  
852 true: {}. You can think of your answer in any way (e.g. step-by-step)  
853 but for the last line of your response, respond only in this format ‘Answer:  
854 <letter 1> <letter 2> <letter 3>’, e.g. ‘Answer: A, B, C’.

855 For models where we first concatenate the input images, the text prompt is:

856 There are two images provided, one on the left and the other on the right.  
857 Which objects are present in both images? Select all choices that are  
858 true: {}. You can think of your answer in any way (e.g. step-by-step)  
859 but for the last line of your response, respond only in this format ‘Answer:  
860 <letter 1> <letter 2> <letter 3>’, e.g. ‘Answer: A, B, C’.

## 861 B Image Taking Guidelines

862 We used the following procedure to guide our creation of images. First, each image taker selected a set  
863 of up to 7 objects and identified a background (e.g. a blanket, counter, or on the floor). Second, they  
864 take images iteratively, starting by placing a single object on the background and subsequently adding  
865 others (N=1 to N=7). Images were framed with the objects in the center or slightly off center (e.g. in  
866 Figure 2b, the plants in the third set of images from the left has leaves outside of the top part of the  
867 frame), with the goal that the majority if not the entirety of the object be contained within the frame.  
868 Across scenes, objects are often viewed from different viewpoints (e.g. top-down, versus side-view).  
869 Objects also may be partially occluded by other objects in the scene (e.g. in the bottom left image in  
870 Figure 2b the eye-mask is slightly occluded by the pink ball), but occlusions should be minimal with  
871 the restriction that all objects be easily human recognizable. For each scene (set of objects against a  
872 background), the image-taker would also take images from multiple visual orientations freely (with  
873 no restriction on the angle between the camera and the objects, so as to better capture real world  
874 diversity). Third, the image-taker would repeat against a new background, and add the objects to  
875 the scene in a different order and at a different orientation. Throughout this process, image-takers  
876 refrained from including any sensitive objects which may have privacy or IP concerns (e.g. humans,  
877 animals, brands, logos etc.) in images. Images were taken using smart phone cameras (Google Pixel,  
878 iPhone 15 Pro), as smart phones are one of the predominant modes of image creation currently.

## 879 C Additional Analysis

880 **Role of Object Similarity** In Table 3, we show the correlation between accuracy and the average  
881 similarity of objects in the scene. We observe a statistically significant negative correlation suggesting  
882 as models are more likely to make mistakes when objects are similar.

883 **Additional model examples and mistakes** In Figure 7, we show additional randomly sampled  
884 examples from Common-0 Bench. In Appendix C, we show randomly selected mistakes in Common-0  
885 Bench across all models. The examples show the high degree to which models hallucination objects  
886 that are not in the ground truth.

## 887 D Synthetic data

888 The synthetic data was generated using Unreal Engine (EpicGames) and assets from Aria  
889 Digital Twins Catalog (Dong et al., 2025). We bought the following asset on fab  
890 to get the floor texture with a professional license: [https://www.fab.com/listings/  
891 66985cc5-13c2-45eb-9b5b-628ef4445a5c](https://www.fab.com/listings/66985cc5-13c2-45eb-9b5b-628ef4445a5c). We randomly placed the assets into one of 16  
892 different positions and apply some slight random rotation over the assets. To ensure that assets are  
893 not overlapping with each other, we constrained them to a given maximum size while keeping their  
894 aspect ratio. For each scene, we took images coming from 4 different camera positions.



Figure 7: Randomly sampled examples from Common-0 Bench.

Model	Pearson Correlation
Qwen 7B	<b>-0.33*</b>
Qwen 32B	<b>-0.38*</b>
Qwen 72B	<b>-0.40*</b>
Llava-OneVision Chat 7B	<b>-0.38*</b>
Llava-OneVision Chat 72B	<b>-0.30*</b>
DeepSeek-VL2 Small	-0.12*
DeepSeek-VL2	<b>-0.30*</b>
LlamaV-o1 11B	-0.29*
LlamaV 3.2 11B	<b>-0.33*</b>
Llama 4 Instruct Scout	<b>-0.41*</b>
PerceptionLM 3B	-0.10
PerceptionLM 8B	<b>-0.35*</b>

Table 3: Correlation between similarity among common objects and accuracy. The negative correlation shows that, the more similar the common objects are lead to lower accuracy. \* indicates statistical significance with correlations of moderate strength or above in bold.

<i>Model</i>	<i>Choices (Enumerated by Letter to Model)</i>	<i>Ground Truth</i>	<i>Prediction</i>
GPT-4o	[silver grater, No objects are in common, dark chocolate bar wrapped in foil, silver straw, silver whisk, silver knife, tangerine, measuring cup]	No objects are in common	Measuring cup
Llava-OneVision	[No objects are in common, mallard (fake duck), vase, hammer, calculator, dish, basketball, fake-foodcan]	B, D, E	D, E, H
Qwen	[dumbbell, mouse, hammer, No objects are in common, football, birdhouse, keyboard, volleyball]	B, D, G	B
PerceptionLM	[spoon, No objects are in common, orange, glass, keys, lime, fork, popcorn kernel]	E	A, B, C
Qwen	[dino, candle holder, mallard (fake duck), bowl, volleyball, No objects are in common, shaver, birdhouse]	D, E, G	D
Llama 4 Instruct	[watermelon, plant, No objects are in common, coffee mug, earbuds, candle snuffer, pen, ball]	D, E, F, H	D
Llama 3.2 Instruct	[bottle opener, gold jigger, 2-prong serving fork, strainer, paring knife with wooden handle, No objects are in common, gold paring knife, silver jigger]	A, B, D, G, H	D, G
Llama 3.2 Instruct	[fakefruit, airplane, bowl, No objects are in common, spoon, football, keyboard, mouse]	C, F, G	D
Llama 3.2 Instruct	[fakefoodcan, vase, volleyball, spoon, kitchenware, No objects are in common, fakefruit, shoes]	A, B, E, G	A, B
Qwen	[remote, basketball, calculator, No objects are in common, mouse, vase, marker, volleyball]	C, E, H	C
Llama 3.2 Instruct	[fish bowl, white pill bottle, paint brush, candy cane, No objects are in common, orange pill bottle, lint roller, scissors]	B, F, H	B, D, F, G, H
PerceptionLM	[No objects are in common, candle, marker, fakefruit, keyboard, mallard (fake duck), bowl, remote]	B, C, E, G, H	A, B, C
GPT-4o	[cup, mallard (fake duck), vase, No objects are in common, football, candle, volleyball, shoes]	candle, shoes, vase, volleyball	shoes, volleyball
Llama 4 Instruct	[spoon, No objects are in common, fakefruit, cast iron, basketball, marker, vase, shoes]	C, D, G, H	C, D, H
Qwen	[spoon, cast iron, basketball, vase, fakefruit, No objects are in common, marker, shoes]	D, E, H	E, H
Qwen	[No objects are in common, fakefoodcan, fakefruit, shoes, spoon, vase, volleyball, kitchenware]	C, F, H	B, C, F
Llava-OneVision	[bowl, keyboard, No objects are in common, marker, remote, fakefruit, candle, mallard (fake duck)]	A, B, D, G	A, B, C
Qwen	[No objects are in common, pail with handle, burnt orange pot, leaf, black pot, easel, pink pot, watering can]	B, C, E, G	A
DeepSeekVL2	[No objects are in common, marker, basketball, calculator, vase, mouse, volleyball, remote]	B, D, E, F, G, H	A, B, C
Llava-OneVision Chat	[black pot, burnt orange pot, pink pot, pail with handle, No objects are in common, leaf, watering can, easel]	A, C	A, D, G

Table 4: Randomly sampled model mistakes in Common-0 Bench.

895 **E Dataset Card**

896 We include a datasheet for Common-0 Bench below, following the example from Gebru et al. (2021).

897 **Motivation**

898 *For what purpose was the dataset created?* The dataset was created to test the reasoning abilities of  
899 multimodal LLMs in multi-image, multi-object settings.

900 *Who created the dataset?* This is redacted during the review process to maintain anonymity and will  
901 be included in the camera-ready.

902 *Who funded the dataset creation?* This is redacted during the review process to maintain anonymity  
903 and will be included in the camera-ready.

904 *Any other comments?* None.

905 **Composition**

906 *What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?*  
907 *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between*  
908 *them; nodes and edges)? Please provide a description.* Each instance is a tuple of 2 images, a set of  
909 potential objects that are in both images and a set of the ground-truth, common objects between both  
910 images.

911 *How many instances are there in total (of each type, if appropriate)?* There are 10586 instances in  
912 Common-0 Bench and 12600 instances in Common-0 Complex.

913 *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances*  
914 *from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative*  
915 *of the larger set (e.g., geographic coverage)?* These were manually created instances, either via the  
916 authors taking the images or the authors using a game engine to synthetically create the images. We  
917 created a large set of synthetic images ( $\approx 400k$ ). For Common-0 Bench ( $N=3$  to  $N=7$  objects) and  
918 Common-0 Complex ( $N=3$  to  $N=7$  objects), we randomly sampled images with the target number  
919 of objects.

920 *Is there a label or target associated with each instance?* The target associated with each instance is  
921 the set of objects in common between both images (e.g. apple, keys).

922 *Is any information missing from individual instances?* All of the information is included for every  
923 instance.

924 *Are relationships between individual instances made explicit (e.g., users' movie ratings, social*  
925 *network links)? If so, please describe how these relationships are made explicit.* Each image  
926 in a given contains a specific configuration of objects. This configuration is taken from multiple  
927 orientations. These orientations are labeled in the data files. Additionally, each image is contained  
928 with multiple instances. The instances in the data file are labeled with the image filenames so it's clear  
929 to see which instances have the same images.

930 *Are there recommended data splits (e.g., training, development/validation, testing)?* This is an  
931 evaluation-only benchmark; we do not provide any training or validation splits.

932 *Are there any errors, sources of noise, or redundancies in the dataset?* The instances were manually  
933 created. Potential sources of noise may come from ambiguity in identifying objects, which is  
934 captured by our human baseline.

935 *Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites,*  
936 *tweets, other datasets)?* The dataset is entirely self-contained.

937 *Does the dataset contain data that might be considered confidential (e.g., data that is protected by*  
938 *legal privilege or by doctor-patient confidentiality, data that includes the content of individuals'*  
939 *nonpublic communications)?* The dataset does not contain any confidential or private information.

940 *Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals*  
941 *race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships,*

942 *or locations; financial or health data; biometric or genetic data; forms of government identification,*  
943 *such as social security numbers; criminal history)?* The dataset does not contain any sensitive  
944 information.

945 *Any other comments?* None.

## 946 **Collection Process**

947 *How was the data associated with each instance acquired?* Every real photo was manually taken by  
948 one of the authors on this paper specifically for this dataset. Every synthetic photo was generated by  
949 the authors using a game engine. We manually wrote the set of objects found in each image.

950 *What mechanisms or procedures were used to collect the data (e.g., hardware apparatuses or sensors,*  
951 *manual human curation, software, programs, software APIs)?* We used manual human curation for  
952 the real images and the Unreal engine for synthetic images. We validated the images by sampling a  
953 subset to hand-annotate.

954 *If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic,*  
955 *probabilistic with specific sampling probabilities)?*

956 For the synthetic images, we manually downsampled via random sampling.

957 *Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and*  
958 *how were they compensated (e.g., how much were crowdworkers paid)?* The authors performed all  
959 components of the data collection. We will include full details about the authors in the camera ready  
960 to preserve anonymity.

961 *Over what timeframe was the data collected?* The data was collected over about 3 months.

962 *Were any ethical review processes conducted (e.g., by an institutional review board)?* The data  
963 collection went through IRB. We did not include humans in the images.

964 *Did you collect the data from the individuals in question directly, or obtain it via third parties or*  
965 *other sources (e.g., websites)?* The data was not collected from external individuals, third parties or  
966 web sources. We manually collected all data.

967 *Were the individuals in question notified about the data collection?* N/A; see previous question.

968 *Did the individuals in question consent to the collection and use of their data?* N/A; see previous  
969 question.

970 *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their*  
971 *consent in the future or for certain uses? If so, please provide a description, as well as a link or other*  
972 *access point to the mechanism (if appropriate).* N/A.

973 *Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data*  
974 *protection impact analysis) been conducted? If so, please provide a description of this analysis,*  
975 *including the outcomes, as well as a link or other access point to any supporting documentation.*  
976 N/A.

977 *Any other comments?* None.

## 978 **Preprocessing/Cleaning/Labeling**

979 *Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokeniza-*  
980 *tion, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing*  
981 *values)? If so, please provide a description. If not, you may skip the remaining questions in this*  
982 *section*

983 We manually collected/generated all dataset instances and therefore did not perform any additional  
984 data processing beyond image resizing. All images in their original size were saved.

985 **Uses**

986 *Has the dataset been used for any tasks already?* The dataset has not been publicly released yet  
987 (outside of the private repository for paper review) and therefore has not been used for any additional  
988 tasks.

989 *Is there a repository that links to any or all papers or systems that use the dataset? If so, please  
990 provide a link or other access point.* The dataset is assessible through Kaggle at this link.

991 *What (other) tasks could the dataset be used for?* Common-0 Bench has been tested for multiple-  
992 choice QA with multiple possible answers. The dataset could also be tested in open-ended question  
993 answering.

994 *Is there anything about the composition of the dataset or the way it was collected and prepro-  
995 cessed/cleaned/labeled that might impact future uses?* There is very minimal risk for harm. We did  
996 not include any pictures of people, real or generated, and we also excluded any logos. Additionally,  
997 this dataset is only for evaluation and therefore will not be used in model training.

998 *Are there tasks for which the dataset should not be used?* The dataset is exclusively for evaluation  
999 and should not be used to train or finetune any models.

1000 *Any other comments?* None.

1001 **Distribution**

1002 *Will the dataset be distributed to third parties outside of the entity (e.g., company, institution,  
1003 organization) on behalf of which the dataset was created? If so, please provide a description.* Yes,  
1004 the dataset will be publicly available on HuggingFace.

1005 *How will the dataset will be distributed (e.g., tarball on website, API, GitHub)? Does the dataset  
1006 have a digital object identifier (DOI)?* We will host the dataset on HuggingFace. Because this paper  
1007 is the introduction of the dataset, we will use the paper DOI.

1008 *When will the dataset be distributed?* The dataset will be distributed upon acceptance of the paper in  
1009 2025.

1010 *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or  
1011 under applicable terms of use (ToU)?* The dataset is being distributed under the non-commercial CC  
1012 BY-NC 4.0 license.

1013 *Have any third parties imposed IP-based or other restrictions on the data associated with the  
1014 instances? If so, please describe these restrictions, and provide a link or other access point to,  
1015 or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these  
1016 restrictions.* No.

1017 *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?  
1018 If so, please describe these restrictions, and provide a link or other access point to, or otherwise  
1019 reproduce, any supporting documentation.* No.

1020 *Any other comments?* None.

1021 **Maintenance**

1022 *Who will be supporting/hosting/maintaining the dataset?* REDACTED AUTHORS will be maintaining  
1023 the dataset.

1024 *How can the owner/curator/manager of the dataset be contacted (e.g., email address)?* REDACTED  
1025 AUTHORS can be contacted through the email addresses provided in the camera ready.

1026 *Is there an erratum? If so, please provide a link or other access point.* There is currently not an  
1027 erratum.

1028 *Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)? If  
1029 so, please describe how often, by whom, and how updates will be communicated to dataset consumers  
1030 (e.g., mailing list, GitHub)?* We will update the dataset for any errors. We will likely communicate  
1031 this via social media and perhaps a GitHub page.

1032 *If the dataset relates to people, are there applicable limits on the retention of the data associated with*  
1033 *the instances (e.g., were the individuals in question told that their data would be retained for a fixed*  
1034 *period of time and then deleted)? If so, please describe these limits and explain how they will be*  
1035 *enforced.* N/A.

1036 *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe*  
1037 *how. If not, please describe how its obsolescence will be communicated to dataset consumers.* N/A

1038 *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to*  
1039 *do so? If so, please provide a description.* We encourage anyone interested in potential augmentations  
1040 and contributions to contact us using our email addresses, listed above.

1041 *Any other comments?* None.