

Common **Inpainted** Objects **In-N-Out** of Context

Tianze Yang* Tyson Jordan* Ruitong Sun* Ninghao Liu Jin Sun
 University of Georgia

{Tianze.Yang, tysonjordan, Ruitong.Sun, ninghao.liu, jinsun}@uga.edu

Abstract

We present *Common Inpainted Objects In-N-Out of Context (COinCO)*, a novel dataset addressing the scarcity of out-of-context examples in existing vision datasets. By systematically replacing objects in COCO images through diffusion-based inpainting, we create 97,722 unique images featuring both contextually coherent and inconsistent scenes, enabling effective context learning. Each inpainted object is meticulously verified and categorized as in- or out-of-context through Large Vision Language Model assessments. We demonstrate three key tasks enabled by COinCO: (1) a fine-grained context reasoning approach that classifies objects as in- or out-of-context based on three criteria; (2) a novel *Objects-from-Context* prediction task that determines which new objects naturally belong in given scenes at both instance and clique level semantics, and (3) context-enhanced fake detection on state-of-the-art methods without fine-tuning. COinCO provides a controlled testbed with contextual variations, establishing a foundation for advancing context-aware visual understanding in computer vision, including image forensics. Code and dataset are available at <https://co-in-co.github.io/>.

1. Introduction

Context is fundamental to visual understanding [3, 13, 43, 45, 61, 66]. When humans view a scene, we instinctively assess the coherence between objects and their environment. This context-based reasoning is essential for judging whether objects naturally belong in a scene [5]. Take a look at Figure 1—do the red-boxed objects seem reasonable in their contexts? Context reasoning can reveal out-of-place objects, providing a complementary cue for detecting manipulated content. By using context, humans assess whether objects appear in plausible settings—a horse in a grassy field with trees in the background appears perfectly natural, while a tiny zebra on a beach next to a person immediately raises suspicion due to its unusual size and location.

*Equal contribution.

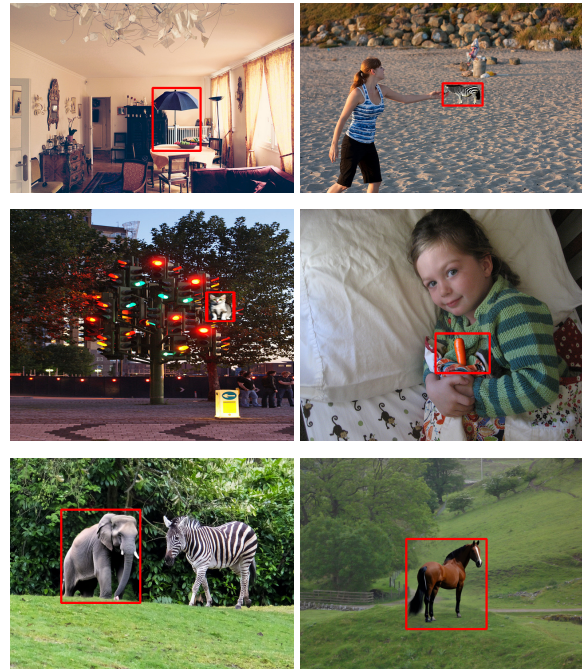


Figure 1. COinCO contains a rich set of inpainted out-of-context objects (top two rows) and in-context objects (bottom row).

Learning context from data, however, is difficult. A significant challenge is that unusual scenes with out-of-context objects are rare in real life. Common computer vision datasets [10, 21, 31, 34] primarily contain objects in their natural settings, resulting in a scarcity of examples with contextual violations. This presents a fundamental obstacle: how can we train machine learning models that require large data to recognize contextual inconsistencies when such examples are inherently uncommon?

To address this challenge, we propose a new dataset with both in-context and out-of-context objects through controlled image manipulation. By systematically replacing objects in existing real scenes, we generate the necessary examples of contextual violations while preserving the overall structure of the scene. We specifically build on top of the Common Objects in Context (COCO) dataset [34]

because it contains a diverse set of everyday scene photographs with rich image- and object-level annotations, offering an ideal starting point for contextual manipulation.

Using a diffusion-based inpainting model [48], we replace exactly one object per COCO image. This selective approach allows us to maintain the broader scene context while introducing precise, controlled variations in object-scene relationships. Our meticulous data creation pipeline and multi-step verification ensure high-quality inpainting results, and we apply a state-of-the-art Large Vision Language Model (LVLM) to classify each inpainted object as either contextually consistent or inconsistent with its scene.

We name our dataset *Common Inpainted Objects In-N-Out of Context (COinCO)*, emphasis on context for inpainted objects. It contains 97,722 inpainted images, each accompanied by rich annotations. Figure 2 shows the pipeline of our data creation and the downstream tasks.

COinCO is the first dataset that features inpainted objects with diverse context annotations, enabling advancements in tasks including fine-grained context classification, object-from-context prediction, and fake detection (as our objects are inpainted by a generative model).

Our main contributions are:

1. We introduce a novel, large-scale dataset of partially manipulated images that enhances COCO, featuring strategically inpainted objects that are either contextually coherent or inconsistent with their scenes.
2. We introduce a fine-grained context reasoning task in COinCO based on three criterion-specific dimensions and train efficient student models through distillation from a 72B LVLM teacher. These models provide modular, interpretable, and high-accuracy context reasoning for inpainted and real images.
3. We introduce a novel task, object-from-context, which aims to predict instance- and clique-level semantic categories for possible objects that fit the given context.
4. We show that context can be effectively integrated into fake detection pipelines, substantially improving state-of-the-art image forensics methods without fine-tuning.

2. Related Work

COCO and its extensions. The Common Objects in Context (COCO) dataset [34] is a widely used benchmark for various computer vision tasks, including object detection, instance segmentation, and captioning. Over the years, several extensions have been proposed. Some datasets primarily focus on *expanding annotations* without modifying image content. COCO-WholeBody refines human keypoint detection by adding facial, hand, and foot keypoints [27], while LVIS introduces a long-tailed distribution with over 1,000 categories for instance segmentation [21]. COCO-Stuff incorporates background (stuff) annotations

for panoptic segmentation [10], and RefCOCO enables referring expression-based object localization [31]. Other works improve *vision-language alignment*, such as COCO-Caption for image captioning [11] and COCO-Text for scene text detection [56]. Additionally, CD-COCO, which applies image distortions to test robustness [4], and COCONut, which unifies multiple segmentation tasks [16], explore *scene complexity*. However, these datasets do not manipulate contextual relationships. Our dataset uniquely reconstructs scene context by *replacing original objects with out-of-context alternatives via inpainting*, creating a testbed for *context modeling* and *fake localization* that enables the learning of complex scene semantics.

Context reasoning. Context is essential for understanding object relationships and finding anomalies in complex scenes [43]. Biederman et al. [5] identified relational principles such as support, probability, and size that help identify contextual inconsistencies, which prior works apply using co-occurrence and support relationships for out-of-context detection [13]. Others have shown that slight context changes can cause significant errors in object detection, and Acharya et al. [1] introduced a Graph Contextual Reasoning Network modeling co-occurrence and relative position to detect out-of-context objects. Context is particularly important for object detection [7, 17]; a comprehensive review is in [58]. Ours is the first work to feature a mixture of in- and out-of-context objects in large-scale data. While context remains under-explored in fake detection, Large Vision Language Models (LVLMs) [15, 28, 35, 36, 54] offer a human-aligned approach to context reasoning; we are the first to leverage them for context-based fake detection.

Fake image generation and detection. Image manipulation has evolved from traditional object insertion [9, 18] to GAN-based synthesis [8, 19, 25, 29, 30] and diffusion models [23, 49], including Stable Diffusion [48], DALL-E [46], and ControlNet [65]. Emerging techniques further tailor diffusion methods for object manipulation, inpainting, and harmonization [12, 52, 64], 3D [37], and relighting [22, 26]. For detection, earlier works classify at image-level [14, 42], while advanced methods localize manipulations: PSCC-Net [39] uses spatio-channel correlation, CAT-Net [32] exploits JPEG artifacts, ManTra-Net [60] captures manipulation traces, and TruFor [20] combines RGB and noise-sensitive fingerprints. For fake image datasets, e.g., GenImage [67], CIFAKE [6], DE-FAKE [49], they focus on fully-synthesized images. Diffusion-based inpainting datasets like COCOglide [20] and TGIF [41] utilize COCO images but they have limited image quantity, object replacements constrained to the same category (thus lacking out-of-context objects), and no context labels. Our dataset contains significant out-of-context fake objects in real scenes with rich characteristics.

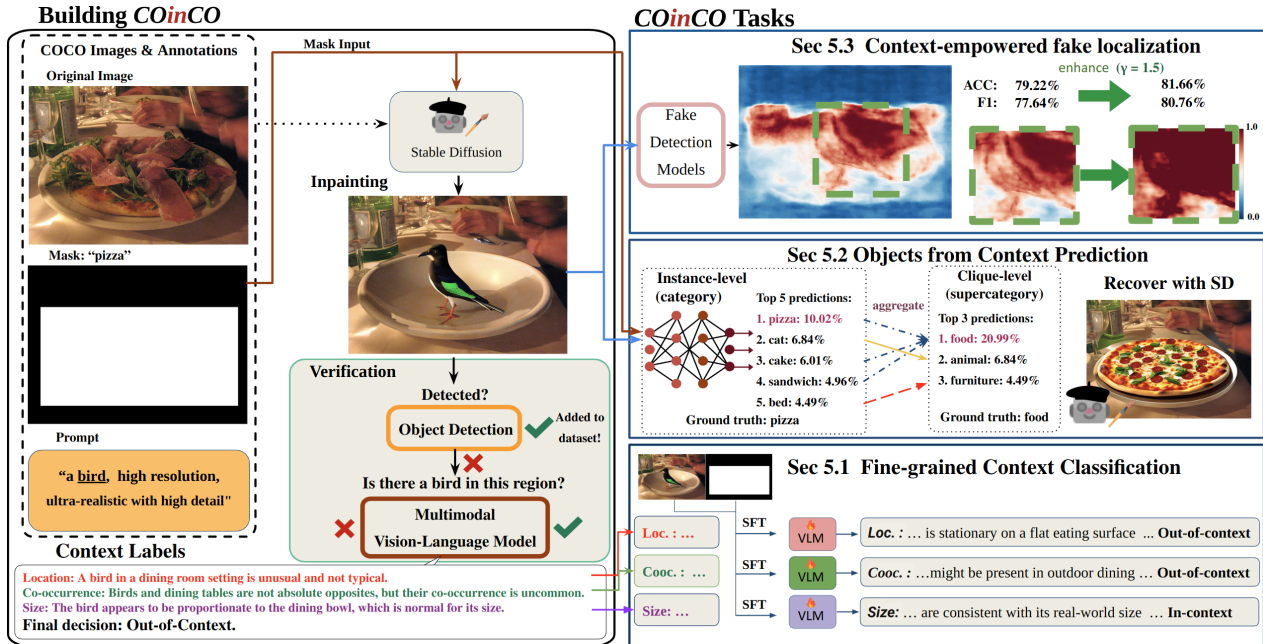


Figure 2. COinCO pipeline. **Left: Building COinCO.** We start with COCO images and annotations. For each image, we randomly select an object and replace it using Stable Diffusion inpainting. We verify the inpainted object using object detection and a vision-language model. Successfully inpainted images are added to the dataset, while failed cases are regenerated and retested. Finally, each inpainted object is labeled as in-context or out-of-context based on three criteria: location, size, and co-occurrence. **Right: COinCO Tasks.** We demonstrate three downstream applications enabled by COinCO. **Sec 5.1:** Fine-grained context classification evaluates objects on each criterion using vision-language models. **Sec 5.2:** Objects-from-context prediction identifies which object categories fit a given context at both instance and clique level semantics, and can recover the original object with an image generation model. **Sec 5.3:** Context-enhanced fake detection improves pretrained fake detectors by enhancing predictions in out-of-context regions.

3. Building COinCO

We propose COinCO, a novel context-oriented inpainted objects dataset derived from the widely-used Common Objects in Context (COCO) Dataset [34]. We specifically choose COCO as our foundation for several reasons. First, we focus on common, everyday objects where contextual reasoning is most meaningful and widely applicable. COCO’s 80 classes span diverse supercategories covering frequent real-world categories such as animals, food, vehicles, and household items, making it the standard benchmark for context-rich scenes. Second, COCO provides reliable instance masks and annotations essential for high-quality inpainting. Expanding to rare or unusual objects would dilute the contextual focus and introduce annotation noise, as contextual expectations become less defined for uncommon categories. Leveraging COCO’s image- and object-level annotations, COinCO enhances COCO by systematically replacing objects with diffusion-based inpainting, providing images with enriched contextual diversity. Our dataset comprises 97,722 successfully inpainted images derived from COCO2017, with each inpainted object categorized as in- or out-of-context with reasoning and accompanied by information about the original object and its

replacement. Task-specific evaluation protocols are detailed in their respective sections.

For each COCO image, we (1) randomly pick an object from the image and inpaint a new object, (2) verify the new object was successfully inpainted with object detection, and (3) perform context reasoning with LVLMs to classify the new object as in- or out-of-context.

3.1. Inpainting

We use Stable Diffusion [48] for inpainting. Inpainting a new object requires an original image, an inpainting mask, and a prompt. For each COCO image, we randomly select an object. The object’s mask is slightly dilated and enclosed in a bounding box as the inpainting mask. Dilation and bounding boxes reduce residual artifacts from the original objects. For replacement, we randomly select one of COCO’s 80 categories as the text prompt.

The vanilla inpainting pipeline often struggles to inpaint small objects. To overcome this, we crop the original image around an enlarged inpainting area, inpaint, and then scale the results back to the original size. We use alpha blending to seamlessly merge the inpainted region with the original image. We found that for stronger inpainting models, such

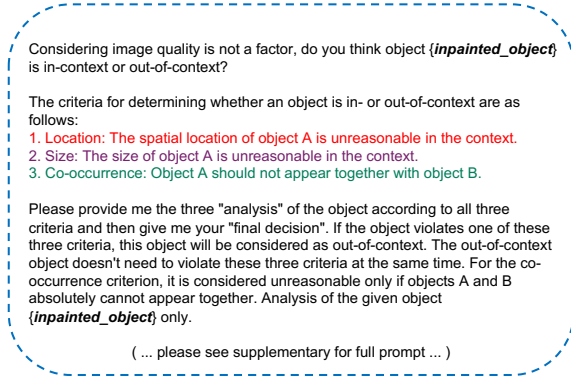


Figure 3. Context reasoning prompt for LVLMs.

as Flux [33] and Qwen Image [59], they focus more on visual fidelity on in-context objects and tend to sacrifice the ability to generate out-of-context objects and provide less precise spatial control over the inpainting region. More details on inpainting and the analysis of different inpainting models are in the supplementary materials.

3.2. Context reasoning

We annotate context labels to all inpainted objects through a multi-model consensus process. Rather than relying on one single large model, three state-of-the-art LVLMs (Molmo-72B [15], Qwen2.5-VL-72B [2], and InternVL3.5-38B [57]) independently evaluate each inpainted image according to three context criteria: location, size, and co-occurrence. We design a structured prompt (Figure 3) to guide the LVLMs in performing this contextual reasoning. For every image, each model first provides criterion-level reasoning by examining (1) whether the spatial placement of the object is appropriate for the scene layout, (2) whether its scale aligns with the scene geometry, and (3) whether its simultaneous presence with existing objects is contextually plausible. An object is labeled as *out-of-context* if it violates any of these criteria; otherwise, it is deemed *in-context*.

After generating criterion-level analyses, each LVM produces a final binary judgment (in-context vs. out-of-context). To ensure the reliability of our context annotations, we retain labels only for images where a unanimous agreement is reached. This consensus requirement yields 73,929 images for which we release both the final in-/out-of-context labels and the complete criterion-level reasoning from all LVLMs. The remaining images are still part of COinCO, but we do not provide context annotations for them due to labeling ambiguity.

3.3. Comparison with prior datasets

COinCO extends COCO by introducing contextual diversity with inpainted objects (Table 1). Unlike COCO and COCO-Stuff, which primarily focus on object and scene

Table 1. COinCO significantly enhances COCO in new ways.

	COCO (2017)	COCO-Stuff	COCOGlide	Ours
Dataset size	123K	164K	512	97,722
Unique source images	123K	164K	512	97,722
Replacement classes	-	-	same as original	80
Out-of-context images	-	-	-	✓
Context reasoning	-	✓	-	✓
Inpainting	-	-	✓	✓
Labels	objects	objects, scenes	(fake) objects	(fake) objects, context reasoning

(stuff) annotations, COinCO provides intentional out-of-context scenarios, enabling research in contextual reasoning and fake localization. While COCOGlide [20] also applies inpainting to COCO, it has significantly less images, and each replacement object has the same class as its original counterpart, thus the context is unaltered.

4. Manual Verification

To validate the reliability of our automated models for object detection and context classification, we conducted a comprehensive manual verification study. We manually annotated 1,000 images to validate the successful rate of our final dataset, and assessed whether successfully inpainted objects were in- or out-of-context from human judgment.

4.1. Verification on inpainted object detection

To select an appropriate detector for our automated inpainting verification pipeline, we evaluated several COCO-trained models on 1,000 human-labeled images where annotators recorded whether the target object was successfully inpainted within the mask region. As shown in Table 2, while GroundingDINO [38] achieved the highest precision (96.14%), its low recall (38.66%) would discard too many successful inpainting results. YOLOv8x [47] achieved the best balance with an F1 score of 78.75%, combining high precision (93.78%) with reasonable recall (67.86%). We therefore selected YOLOv8x as our detector and use it to verify inpainted objects: if YOLOv8x detects the target object within the inpainting mask, the process is deemed successful; for failed cases, we complete up to two additional rounds of inpainting and verification; images that fail all three rounds are discarded, resulting in 97,722 successful images. Notably, false positives often correspond to lower-quality inpaintings where the target object is present but poorly rendered, rather than complete inpainting failures.

4.2. Verification on context classification

Following the procedure in Section 3.2, we retained 73,929 images where all three LVLMs reached unanimous agreement on the final decision, with 64,879 labeled as out-of-context and 9,050 as in-context. To validate the quality of these context annotations, we randomly sampled 1,000 im-

Table 2. Object detection performance on inpainted objects.

Metric	YOLOv8x	YOLOv11x	MMDet	GroundDINO
Precision	93.78	94.80	86.87	96.14
Recall	67.86	50.93	58.54	38.66
F1	78.75	66.26	69.95	55.14

ages (300 in-context, 700 out-of-context) for human verification. Human annotators classified each image using the same context criteria. The LVLMS consensus achieved 95.10% agreement with human annotations, confirming the high reliability of our labels.

Among the 1,000 annotated images, we identified 164 cases where the three LVLMS agreed on the final out-of-context decision but disagreed on at least one specific criterion. Human evaluation of these disagreements revealed that Qwen2.5VL-72B achieved the highest accuracy at 87.80% (144/164). For the remaining 836 images where all LVLMS agreed on both the final decision and all criteria, human annotators assessed reasoning quality. Qwen2.5VL-72B received 94.70% preference for its detailed and interpretable explanations. Based on these validation results, we adopt Qwen2.5VL-72B’s reasoning as the primary annotations for all 73,929 images in our dataset, with the consensus from MolMo and InternVL serving to ensure annotation reliability through multi-model verification.

5. COinCO-enabled Context Tasks

With COinCO as our testbed, we explore the role of context in several tasks. We demonstrate how our data can be used to train efficient context classifiers through knowledge distillation, fine-tuning smaller models to evaluate objects based on our three context criteria. We then propose a novel Objects-from-Context task for predicting which objects naturally belong in a given scene. Finally, we evaluate state-of-the-art fake detection models and show that context can enhance fake localization without any fine-tuning. These applications demonstrate the versatility of COinCO and underscore the fundamental role of context in visual understanding. Examples of those tasks are in Figure 4.

5.1. Fine-grained context classification

Our context classification task addresses a fundamental question: “Does an object belong in this context?” To answer this question effectively at scale, we need models that can perform accurate context reasoning while being practical for deployment. While our verification showed that Qwen2.5-VL-72B achieved excellent performance in context classification, its large size makes it impractical for many real-world deployment scenarios.

To address this challenge, we employ knowledge distillation [24, 35, 53, 63], a proven technique that trans-

Table 3. Context reasoning accuracy (%) on COinCO test set and original COCO images. Results on original COCO confirm models learn genuine context reasoning rather than inpainting artifacts.

Criterion	COinCO Test Set			Original COCO Images		
	Before	After	Δ	Before	After	Δ
Size	55.4	65.9	+10.5	52.6	83.2	+30.6
Location	67.6	75.4	+7.8	47.6	91.4	+43.8
Co-occurrence	75.5	79.9	+4.4	89.0	87.0	-2.0
Average	66.2	73.7	+7.6	63.1	87.2	+24.1

fers knowledge from large teacher models to smaller, more efficient student models. We use Qwen2.5-VL-72B as the teacher to distill its knowledge into three independent Qwen2.5-VL-3B student models—one for location, one for size, and one for co-occurrence reasoning. This distillation process reduces the model size by 24× while maintaining most of the reasoning capabilities, making context classification feasible for practical deployment. Note that this knowledge distillation task is possible thanks to the rich context reasoning labels in COinCO.

We use the 73,929 unanimous consensus images from COinCO to create balanced evaluation splits for training our distilled models. Specifically, we sample 24,000 images maintaining a 1:1 ratio of in-context to out-of-context examples, which we divide into 20,000 for training, 2,000 for validation, and 2,000 for testing. This balanced sampling ensures unbiased evaluation of context reasoning capabilities. We use context reasoning responses from QwenVL as the ground truth (discussed in Section 4.2). Importantly, each student model learns not only the final decision (in-context or out-of-context) but also the reasoning process for its specific criterion, enabling interpretable explanations.

Table 3 shows consistent improvements for each student model across all criteria. On the COinCO test set, size reasoning improves most (+10.5% accuracy), while co-occurrence, despite the strongest baseline, reaches nearly 80% accuracy, confirming successful knowledge transfer from the 72B teacher to 3B students.

Validation on original COCO images. To verify that our models learn genuine semantic relationships and context rather than fitting to inpainting artifacts, we test them on 2,000 original COCO images without any manipulation. The ground-truth values of all three criteria are true as the objects are considered as in-context. The co-occurrence model achieves 89.0% accuracy on original images before fine-tuning, demonstrating it captured natural object relationships. After fine-tuning on COinCO, accuracy slightly decreases to 87.0%, suggesting the model already learned robust co-occurrence patterns that need no further adjustment. The location model shows the largest improvement after fine-tuning (+43.8% accuracy), indicating that exposure to diverse object placements in COinCO helps the

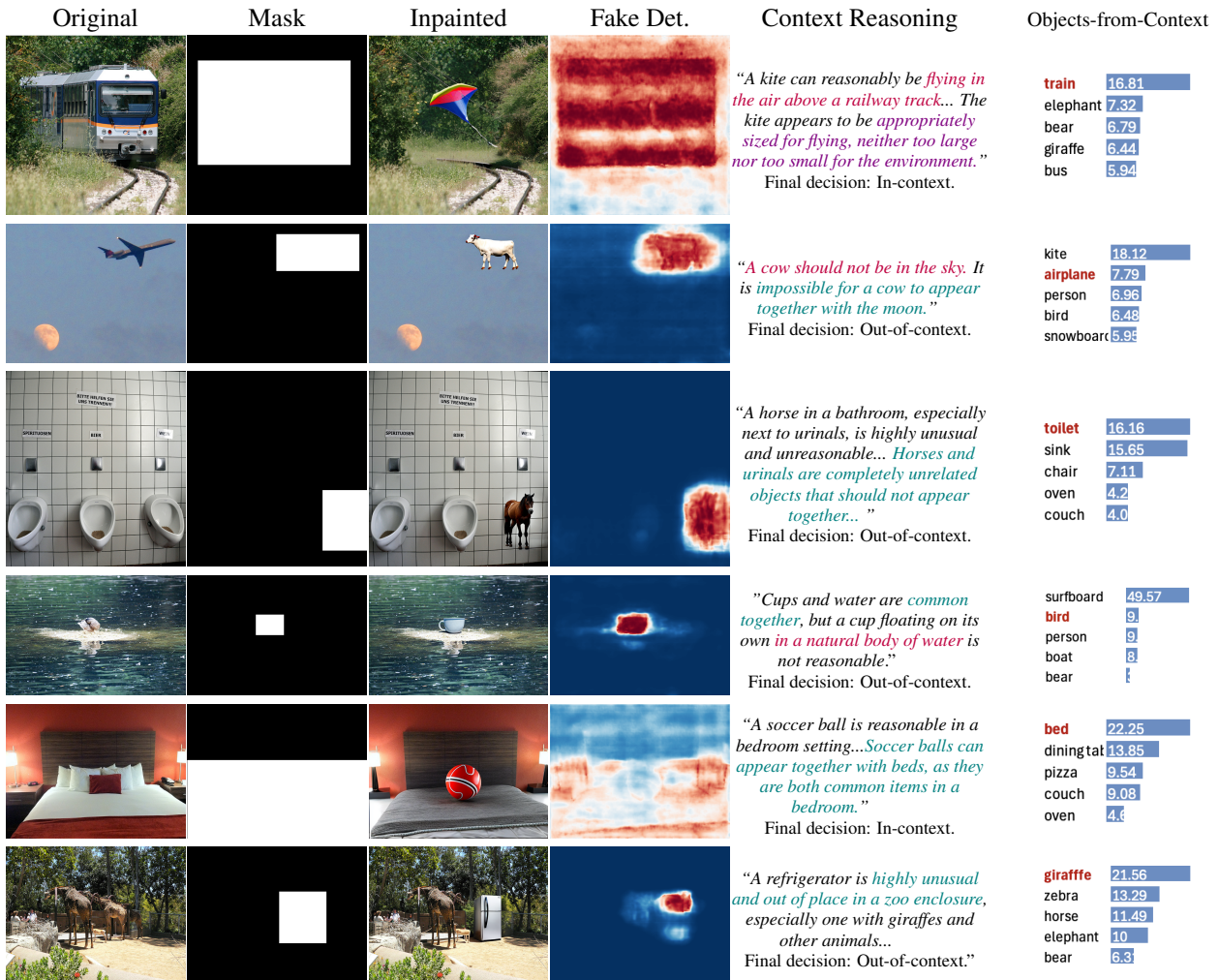


Figure 4. Inpainting, fake detection, and objects-from-context results. Context reasoning responses are color-coded by location, size, and co-occurrence. Original objects are in red. Inpainted objects: kite, cow, horse, cup, sports ball, refrigerator.

model learn spatial reasoning that transfers to real images. The size model also improves substantially (+30.6% accuracy), learning appropriate scale relationships. For location and size reasoning, models achieve an average improvement of +37.2% accuracy on original COCO images, with the high performance on unmanipulated images confirming these models understand genuine contextual principles rather than memorizing dataset-specific patterns.

Each model provides interpretable natural language explanations for its decisions. Through distillation, these 3B models achieve practical inference speeds while maintaining reasoning quality [24, 62]. The modular design allows criterion-specific refinements, demonstrating that complex contextual understanding can be decomposed into focused reasoning tasks. While the 72B teacher model is too slow for practical deployment, these efficient 3B student models make context reasoning feasible for real-world applications.

5.2. Objects-from-Context prediction

While context classification models determine if an existing object belongs in a scene, our Objects-from-Context prediction tackles a different question: "What object(s) fit this context?" This task tests a model's contextual comprehension of real images and aligns with the Context Challenge in [55]. We formulate this task at two levels: 1) **Instance-level prediction**. Predicting the object class among COCO's 80 classes. 2) **Clique-level prediction**. Predicting the clique an object belongs to, using COCO's super-categories: accessory, animal, appliance, electronic, food, furniture, indoor, kitchen, outdoor, person, sports, and vehicle.

We design a model that takes two inputs: the inpainted image and a binary mask indicating the target region. During training, we use the dilated bounding box of the original object as the mask and its class as the label. We use the



Instance		Clique		Instance		Clique		Instance		Clique	
skis	31.6	sports	79.2	kite	22.9	sports	75.1	horse	5.0	animal	37.2
snowboard	24.3	accessory	6.6	skis	22.1	accessory	6.1	zebra	5.0	vehicle	14.9
kite	14.8	person	4.9	snowboard	19.6	vehicle	5.7	giraffe	4.9	sports	12.1
surfboard	7.0	vehicle	4.2	surfboard	8.6	person	5.5	person	4.6	outdoor	10.7
backpack	5.9	animal	3.0	person	5.5	animal	4.2	elephant	4.4	accessory	6.8

Figure 5. Object-from-Context prediction. A red box in each image indicates the query region. The top row shows three examples (two inpainted, one original COCO). The bottom row lists instance-level and clique-level predictions with their probabilities (P(%)). Objects in red are the top predictions.

Table 4. Instance-level and clique-level object-from-context prediction accuracy.

Method	Instance-level (%)			Clique-level (%)		
	Top-1	Top-3	Top-5	Top-1	Top-3	Top-5
Random	1.25	3.75	6.25	8.33	25.00	41.67
Co-occurrence [40]	1.54	4.70	7.29	9.37	30.72	52.91
Ours	16.32	31.89	42.80	35.10	61.41	78.31

VAE encoder from Stable Diffusion to extract latents from both inputs, which are processed by an MLP that predicts across COCO’s 80 classes. For clique-level evaluation, we map predicted object classes to their corresponding super-categories and check if they match the original objects. Importantly, our model can handle arbitrary mask sizes and locations during inference, making it adaptable for broader context reasoning applications. Detailed model architectures and training information are in the supplementary material. For comparison, we implement a baseline that ranks candidate objects based on the co-occurrence frequency of other objects in a scene [40]. If no other objects are present, this baseline defaults to random selection. We evaluate our Objects-from-Context model on the 2,402 images inpainted from the COCO2017 validation set.

Table 4 shows our model’s superior grasp of contextual features, significantly outperforming random and co-occurrence baselines. At the instance level, the substantial improvement in top-3 and top-5 accuracy (31.89% and 42.80%) indicates that even when the model’s top prediction is incorrect, the true object is often ranked among the most probable candidates. Our model demonstrates stronger performance at the clique level, correctly predicting semantically similar objects. This result demonstrates

Table 5. SOTA fake detection performance on COinCO.

Method	Acc	F1	AUC	AP
ManTraNet [60]	85.4	30.7	84.9	50.7
Trufor [20]	89.7	55.9	93.4	73.6
PSCC-Net [39]	89.5	46.8	95.4	79.5
CAT-Net [32]	92.7	76.5	97.4	90.3

that our approach effectively learns the relationship between scene contexts and the objects that belong within them. Our model offers three key applications as shown in Figure 5: (1) analyzing suspicious regions to predict what objects existed before manipulation, serving as a reference for anomaly detection and forensic analysis [44]; (2) suggesting contextually appropriate objects based on spatial location within a scene [50, 51], enabling intelligent image editing [12]; and (3) versatile analysis of any image (including unmodified images) to identify contextually coherent objects in regions of interest, supporting context-aware content generation. This versatility advances context-driven visual understanding across multiple practical applications.

5.3. Context-empowered fake localization

Fake localization aims to identify the specific synthetic regions in an image, producing a pixel-level map that is fine-grained compared to binary image-level classification. In COinCO, the ground truth mask for fake localization is the original object’s entire bounding box used in inpainting. We evaluate localization performance on the 2,402 test images derived from COCO2017 validation using four common pixel-level metrics [32, 39]: F1 Score and Accuracy (threshold of 0.5), Area Under Curve (AUC), and Average

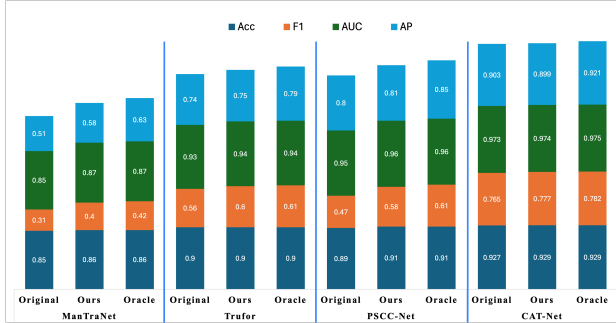


Figure 6. Context enhancement results.

Precision (AP).

State-of-the-art performance. We benchmark several SOTA fake detection models on COinCO (Table 5). The high accuracy and AUC scores ($> 84\%$ for all models) can be attributed to the small average size of inpainted regions in our data. As most images are majority authentic, these metrics are biased towards high values. F1 and AP, which better reflect the precision of fake localization, reveal a clear performance gap among models. CAT-Net achieves the best performance across all metrics with 92.65% accuracy, 76.47% F1 score, 97.35% AUC, and 90.33% AP. Trufor obtains the second-best F1 score of 55.92%, while PSCC-Net shows strong performance in AUC (95.38%) and AP (79.52%). ManTraNet demonstrates relatively lower performance in F1 score (30.68%) and AP (50.72%). These findings highlight significant differences and room for improvement among SOTA fake localizers. Examples of these SOTA models’ predictions are shown in Figure 7.

Incorporating context in fake localization. To leverage contextual information for improved fake localization, we propose a simple yet effective way of enhancing the prediction scores of fake localization models in regions corresponding to out-of-context objects. Concretely, the context-enhanced prediction score $P'(x, y)$ at pixel (x, y) is defined as $\min(P(x, y) \times \gamma, 1.0)$, if $(x, y) \in M_{OC}$; otherwise as $P(x, y)$, the original predicted fake score, M_{OC} is the mask region of out-of-context object (predicted by context reasoning), γ is a context enhancement factor, and the min function ensures the enhanced score ≤ 1.0 .

We evaluate this context-enhancement approach under two settings. In the **oracle** setting, we use ground truth annotations to enhance predictions within the fake object’s mask region when it is out-of-context. This serves as an upper bound for performance gains. In the **practical** setting, the fake objects are unknown. We propose the use of Molmo-7B [15] to identify suspicious objects based on *size* and *location*, excluding *co-occurrence* to prevent false positives. For instance, if an image contains only two objects, e.g., an apple and an inpainted traffic light, either might be seen as out-of-context. By focusing on physical violations,

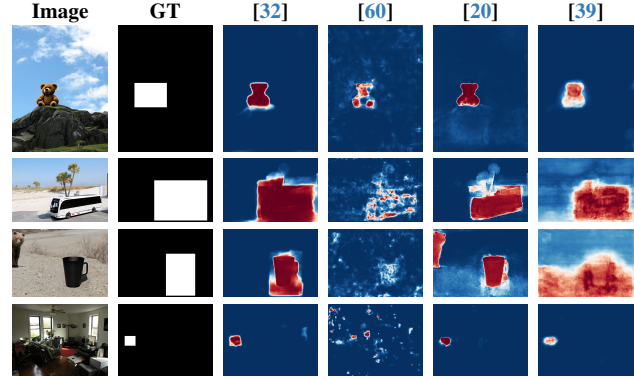


Figure 7. Fake localization performance.

our method remains robust while effectively detecting fake objects. As Molmo may identify multiple out-of-context objects in an image, we define M_{OC} as the union of these objects’ masks. In both settings, we set the enhancement factor γ to 5 to enhance prediction scores in out-of-context regions. The analysis of γ is in supplementary materials.

Figure 6 shows that our context-enhancement improves fake localization for SOTA detectors *without any finetuning*. The effectiveness of our method stems from the complementary nature of context detection and fake localization. Thus, even if LVLm mistakes authentic objects as out-of-context, enhancement only strengthens the base model’s predictions. Notably, out-of-context objects are not rare in real-world forgeries, especially in partial manipulation scenarios, where inserted objects often exhibit semantic inconsistencies in size, location, or scene compatibility. Our results highlight that context is valuable for fake detection.

6. Conclusion

We present COinCO, a large-scale dataset featuring both in-context and out-of-context objects created through controlled diffusion-based inpainting. By strategically replacing COCO objects, COinCO introduces diverse contextual configurations and provides rich criterion-level annotations generated through multiple LVLms consensus. These annotations enable training efficient, interpretable context classifiers through knowledge distillation, and support a new Objects-from-Context prediction task that models what objects naturally belong in a scene. We further show that contextual cues serve as a powerful complementary signal for fake localization, improving SOTA detectors without any fine-tuning. Overall, COinCO establishes a comprehensive testbed for advancing context-aware visual understanding, contextual reasoning, and image forensics. Future work could extend context reasoning to open-vocabulary settings beyond COCO’s predefined categories, enabling more generalizable contextual understanding.

7. Acknowledgments

This work was funded in part by the National Science Foundation (SES-2521631) and the University of Georgia’s Presidential Interdisciplinary Seed Grant.

References

- [1] Manoj Acharya, Anirban Roy, Kaushik Koneripalli, Susmit Jha, Christopher Kanan, and Ajay Divakaran. Detecting out-of-context objects using graph context reasoning network. In *IJCAI*, 2022. 2
- [2] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 4
- [3] Moshe Bar. Visual objects in context. *Nature Reviews Neuroscience*, 5(8):617–629, 2004. 1
- [4] Ayman Beghdadi, Azeddine Beghdadi, Malik Malle, Lotfi Beji, and Faouzi Alaya Cheikh. Cd-coco: A versatile complex distorted coco database for scene-context-aware computer vision. In *2023 11th European Workshop on Visual Information Processing (EUVIP)*, pages 1–6. IEEE, 2023. 2
- [5] Irving Biederman, Robert J Mezzanotte, and Jan C Rabinowitz. Scene perception: Detecting and judging objects undergoing relational violations. *Cognitive psychology*, 14(2):143–177, 1982. 1, 2
- [6] Jordan J. Bird and Ahmad Lotfi. Cifake: Image classification and explainable identification of ai-generated synthetic images, 2023. 2
- [7] Philipp Bomatter, Mengmi Zhang, Dimitar Karev, Spandan Madan, Claire Tseng, and Gabriel Kreiman. When pigs fly: Contextual reasoning in synthetic and natural scenes. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 255–264, 2021. 2
- [8] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale gan training for high fidelity natural image synthesis. *arXiv preprint arXiv:1809.11096*, 2018. 2
- [9] Aurélie Bugeau, Marcelo Bertalmío, Vicent Caselles, and Guillermo Sapiro. A comprehensive framework for image inpainting. *IEEE transactions on image processing*, 19(10):2634–2645, 2010. 2
- [10] Holger Caesar, Jasper Uijlings, and Vittorio Ferrari. Coco-stuff: Thing and stuff classes in context. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1209–1218, 2018. 1, 2
- [11] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. 2
- [12] Xi Chen, Lianghua Huang, Yu Liu, Yujun Shen, Deli Zhao, and Hengshuang Zhao. Anydoor: Zero-shot object-level image customization, 2024. 2, 7
- [13] Myung Jin Choi, Antonio Torralba, and Alan S Willsky. Context models and out-of-context objects. *Pattern Recognition Letters*, 33(7):853–862, 2012. 1, 2
- [14] L Minh Dang, Syed Ibrahim Hassan, Suhyeon Im, Jaecheol Lee, Sujin Lee, and Hyeonjoon Moon. Deep learning based computer generated face identification using convolutional neural network. *Applied Sciences*, 8(12):2610, 2018. 2
- [15] Matt Deitke, Christopher Clark, Sangho Lee, Rohun Tripathi, Yue Yang, Jae Sung Park, Mohammadreza Salehi, Niklas Muennighoff, Kyle Lo, Luca Soldaini, et al. Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 91–104, 2025. 2, 4, 8
- [16] Xueqing Deng, Qihang Yu, Peng Wang, Xiaohui Shen, and Liang-Chieh Chen. Coconut: Modernizing coco segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21863–21873, 2024. 2
- [17] Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *Proceedings of the european conference on computer vision (ECCV)*, pages 364–380, 2018. 2
- [18] Hany Farid. Image forgery detection. *IEEE Signal processing magazine*, 26(2):16–25, 2009. 2
- [19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, 2020. 2
- [20] Fabrizio Guillaro, Davide Cozzolino, Avneesh Sud, Nicholas Dufour, and Luisa Verdoliva. Trufor: Leveraging all-round clues for trustworthy image forgery detection and localization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 20606–20615, 2023. 2, 4, 7, 8
- [21] Agrim Gupta, Piotr Dollar, and Ross Girshick. Lvis: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5356–5364, 2019. 1, 2
- [22] Mingming He, Pascal Clausen, Ahmet Levent Taşel, Li Ma, Oliver Pilarski, Wenqi Xian, Laszlo Rikker, Xueming Yu, Ryan Burgert, Ning Yu, et al. Diffrelight: Diffusion-based facial performance relighting. *arXiv preprint arXiv:2410.08188*, 2024. 2
- [23] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020. 2
- [24] Cheng-Yu Hsieh, Chun-Liang Li, Chih-Kuan Yeh, Hootan Nakhost, Yasuhisa Fujii, Alex Ratner, Ranjay Krishna, Chen-Yu Lee, and Tomas Pfister. Distilling step-by-step! outperforming larger language models with less training data and smaller model sizes. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8003–8017, 2023. 5, 6
- [25] Phillip Isola, Jun-Yan Zhu, Tinghui Zhou, and Alexei A Efros. Image-to-image translation with conditional adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1125–1134, 2017. 2
- [26] Haian Jin, Yuan Li, Fujun Luan, Yuanbo Xiangli, Sai Bi, Kai Zhang, Zexiang Xu, Jin Sun, and Noah Snavely. Neural gaffer: Relighting any object via diffusion, 2024. 2

- [27] Sheng Jin, Lumin Xu, Jin Xu, Can Wang, Wentao Liu, Chen Qian, Wanli Ouyang, and Ping Luo. Whole-body human pose estimation in the wild. In *European Conference on Computer Vision*, pages 196–214. Springer, 2020. 2
- [28] Yang Jin, Kun Xu, Liwei Chen, Chao Liao, Jianchao Tan, Quzhe Huang, Bin Chen, Chenyi Lei, An Liu, Chengru Song, et al. Unified language-vision pretraining in llm with dynamic discrete visual tokenization. *arXiv preprint arXiv:2309.04669*, 2023. 2
- [29] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 2
- [30] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2
- [31] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. Referitgame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 787–798, 2014. 1, 2
- [32] Myung-Joon Kwon, Seung-Hun Nam, In-Jae Yu, Heung-Kyu Lee, and Changick Kim. Learning jpeg compression artifacts for image manipulation detection and localization. *International Journal of Computer Vision*, 130(8):1875–1895, 2022. 2, 7, 8
- [33] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. 4
- [34] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision—ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6–12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 2, 3
- [35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 2, 5
- [36] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2
- [37] Ruoshi Liu, Rundi Wu, Basile Van Hoorick, Pavel Tokmakov, Sergey Zakharov, and Carl Vondrick. Zero-1-to-3: Zero-shot one image to 3d object, 2023. 2
- [38] Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, et al. Grounding dino: Marrying dino with grounded pre-training for open-set object detection. In *European conference on computer vision*, pages 38–55. Springer, 2024. 4
- [39] Xiaohong Liu, Yaojie Liu, Jun Chen, and Xiaoming Liu. Psc-net: Progressive spatio-channel correlation network for image manipulation detection and localization. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(11):7505–7517, 2022. 2, 7, 8
- [40] Stephen C Mack and Miguel P Eckstein. Object co-occurrence serves as a contextual cue to guide and facilitate visual search in a natural viewing environment. *Journal of vision*, 11(9):9–9, 2011. 7
- [41] Hannes Mareen, Dimitrios Karageorgiou, Glenn Van Wallelael, Peter Lambert, and Symeon Papadopoulos. Tgif: Text-guided inpainting forgery dataset. In *2024 IEEE International Workshop on Information Forensics and Security (WIFS)*, pages 1–6. IEEE, 2024. 2
- [42] Huaxiao Mo, Bolin Chen, and Weiqi Luo. Fake faces identification via convolutional neural network. In *Proceedings of the 6th ACM workshop on information hiding and multimedia security*, pages 43–47, 2018. 2
- [43] Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in cognitive sciences*, 11(12):520–527, 2007. 1, 2
- [44] Guansong Pang, Chunhua Shen, Longbing Cao, and Anton Van Den Hengel. Deep learning for anomaly detection: A review. *ACM computing surveys (CSUR)*, 54(2):1–38, 2021. 7
- [45] Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *2007 IEEE 11th international conference on computer vision*, pages 1–8. IEEE, 2007. 1
- [46] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 2
- [47] Joseph Redmon, Santosh Divvala, Ross Girshick, and Ali Farhadi. You only look once: Unified, real-time object detection. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 779–788, 2016. 4
- [48] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10684–10695, 2022. 2, 3
- [49] Zeyang Sha, Zheng Li, Ning Yu, and Yang Zhang. De-fake: Detection and attribution of fake images generated by text-to-image generation models. In *Proceedings of the 2023 ACM SIGSAC Conference on Computer and Communications Security*, pages 3418–3432, 2023. 2
- [50] Jin Sun and David W Jacobs. Seeing what is not there: Learning context to determine where objects are missing. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5716–5724, 2017. 7
- [51] Jin Sun, Hadar Averbuch-Elor, Qianqian Wang, and Noah Snavely. Hidden footprints: Learning contextual walkability from 3d human trails. In *European Conference on Computer Vision*, pages 192–207. Springer, 2020. 7
- [52] Ruitong Sun, Tianze Yang, Wei Niu, and Jin Sun. Ousac: Optimized guidance scheduling with adaptive caching for dit acceleration. *arXiv preprint arXiv:2512.14096*, 2025. 2

- [53] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. Stanford alpaca: An instruction-following llama model. https://github.com/tatsu-lab/stanford_alpaca, 2023. 5
- [54] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023. 2
- [55] Antonio Torralba. Contextual priming for object detection. *International journal of computer vision*, 53(2):169–191, 2003. 6
- [56] Andreas Veit, Tomas Matera, Lukas Neumann, Jiri Matas, and Serge Belongie. Coco-text: Dataset and benchmark for text detection and recognition in natural images. *arXiv preprint arXiv:1601.07140*, 2016. 2
- [57] Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025. 4
- [58] Xuan Wang and Zhigang Zhu. Context understanding in computer vision: A survey. *Computer Vision and Image Understanding*, 229:103646, 2023. 2
- [59] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Sheng ming Yin, Shuai Bai, Xiao Xu, Yilei Chen, Yuxiang Chen, Zecheng Tang, Zekai Zhang, Zhengyi Wang, An Yang, Bowen Yu, Chen Cheng, Dayiheng Liu, Deqing Li, Hang Zhang, Hao Meng, Hu Wei, Jingyuan Ni, Kai Chen, Kuan Cao, Liang Peng, Lin Qu, Minggang Wu, Peng Wang, Shuting Yu, Tingkun Wen, Wensen Feng, Xiaoxiao Xu, Yi Wang, Yichang Zhang, Yongqiang Zhu, Yujia Wu, Yuxuan Cai, and Zenan Liu. Qwen-image technical report, 2025. 4
- [60] Yue Wu, Wael AbdAlmageed, and Premkumar Natarajan. Mantra-net: Manipulation tracing network for detection and localization of image forgeries with anomalous features. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9543–9552, 2019. 2, 7, 8
- [61] Kai Xiao, Logan Engstrom, Andrew Ilyas, and Aleksander Madry. Noise or signal: The role of image backgrounds in object recognition. *arXiv preprint arXiv:2006.09994*, 2020. 1
- [62] Xiaohan Xu, Ming Li, Chongyang Tao, Tao Shen, Reynold Cheng, Jinyang Li, Can Xu, Dacheng Tao, and Tianyi Zhou. A survey on knowledge distillation of large language models, 2024. 6
- [63] Tianze Yang, Yucheng Shi, Mengnan Du, Xuansheng Wu, Qiaoyu Tan, Jin Sun, and Ninghao Liu. Concept-centric token interpretation for vector-quantized generative models. In *International Conference on Machine Learning*, pages 71034–71050. PMLR, 2025. 5
- [64] Bo Zhang, Yuxuan Duan, Jun Lan, Yan Hong, Huijia Zhu, Weiqiang Wang, and Li Niu. Controlcom: Controllable image composition using diffusion model, 2023. 2
- [65] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Adding conditional control to text-to-image diffusion models. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 3836–3847, 2023. 2
- [66] Mengmi Zhang, Claire Tseng, and Gabriel Kreiman. Putting visual object recognition in context. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12985–12994, 2020. 1
- [67] Mingjian Zhu, Hanting Chen, Qiangyu Yan, Xudong Huang, Guanyu Lin, Wei Li, Zhijun Tu, Hailin Hu, Jie Hu, and Yunhe Wang. Genimage: A million-scale benchmark for detecting ai-generated image, 2023. 2