

Gen-n-Val: Agentic Image Data Generation and Validation

Supplementary Material

Overview

This supplementary material extends the main manuscript with implementation details, expanded analyses, and qualitative examples: (i) additional results and analyses (Sec. A); (ii) discussions of limitations (Sec. B) and societal impacts (Sec. C); (iii) full implementation details of the Gen-n-Val pipeline (prompt generation, instance generation, filtering, optional background synthesis, augmentation, and training setup) (Sec. D); (iv) qualitative comparisons of baseline and Gen-n-Val outputs (Sec. E); (v) prompt optimization comparisons of initial and optimized prompts (Sec. F); (vi) foreground image filtering examples (Sec. G); and (vii) synthetic data examples for COCO and LVIS (Sec. H).

A. Additional Results and Analyses

Table S.1. **Ablation study of Mask R-CNN [16] on the LVIS [14] benchmark.** Comparison of performance under different numbers of pasted instances.

# of Pasted Instances	mAP ^{box}	mAP _r ^{box}	mAP ^{mask}	mAP _r ^{mask}
1	24.7	13.0	23.9	14.7
3	25.4	13.3	24.1	13.6
5	25.5	14.9	26.6	15.5

A.1. Number of Pasted Instances

We evaluate the performance of Mask R-CNN [16] on the LVIS benchmark [14] using synthetic training data generated with different numbers of pasted instances per image. As shown in Table S.1, pasting 5 instances per image yields the best performance.

A.2. Synthetic-Only and Instance-Count-Matched

To further analyze whether the gains come from the quality of synthetic data rather than simply seeing more training instances, we conduct a controlled study on LVIS rare categories using Mask R-CNN [16]. Specifically, we compare the same number of real and synthetic training instances, and also evaluate an instance-count-matched setting.

As shown in Table S.2, with 4k training instances, Gen-n-Val synthetic data already performs comparably to or better than 4k real LVIS instances. Moreover, under the same total number of instances, *4k real + 4k synthetic* outperforms *8k real*, suggesting that the gains are not merely due to seeing more data, but also due to the usefulness of the generated instances for long-tail categories.

Table S.2. **Controlled study on LVIS rare categories with Mask R-CNN [16].** Comparison between real and Gen-n-Val synthetic training instances.

Training Data	mAP ^{box}	mAP ^{mask}
4k Real (LVIS)	1.78	1.32
4k Synthetic (Gen-n-Val)	1.82	1.78
8k Real	5.06	5.52
4k Real + 4k Synthetic (Gen-n-Val)	6.28	6.64

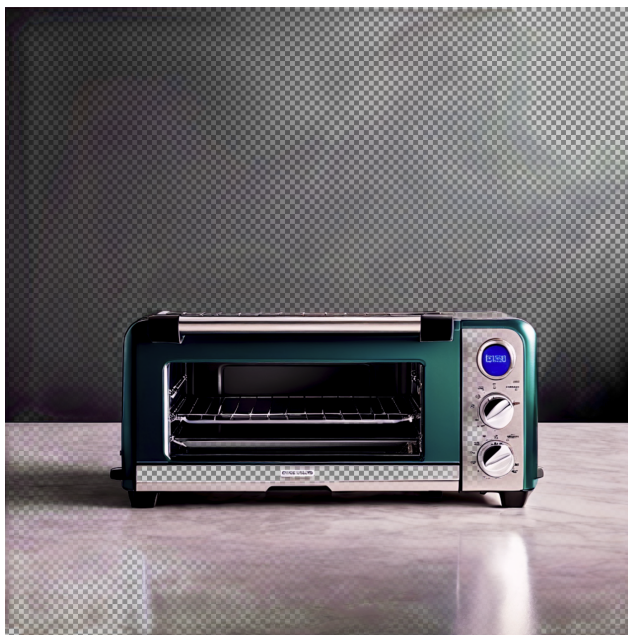


Figure S.1. **Semi-transparent background residue.** A representative false-acceptance case caused by imperfect layer separation.

In addition, all LVIS experiments in our paper are iteration-based and use the same number of training iterations for the baseline (real-only) and Gen-n-Val (real+synthetic). Therefore, Gen-n-Val does not receive extra optimization steps in the LVIS setting.

A.3. Validation Reliability and Failure Analysis

We manually audited 300 accepted and 300 rejected foreground instances from the validation agent. We observe 5.7% (17/300) false acceptances and 32.0% (96/300) false rejections. The dominant false-acceptance failure mode is imperfect layer separation, where a small amount of semi-transparent background residue remains around the generated foreground. An example is shown in Fig. S.1.



Figure S.2. **Conservative validator rejection example.** The VLLM rejects this egg instance because the cracked shell violates the “intact object” constraint, illustrating that many false rejections come from over-strict filtering.

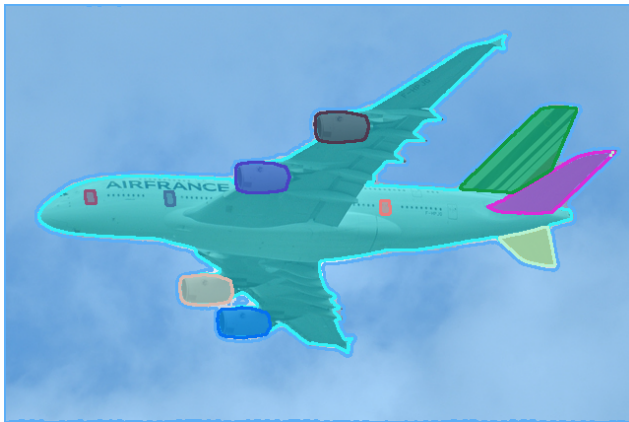


Figure S.3. **SAM may split a single object into multiple masks.** This fragmentation makes the extracted foreground less suitable for direct synthetic instance generation.

Most false rejections are caused by overly strict validation rather than unusable samples. For example, in Fig. S.2, the validator rejects an egg instance because it considers the cracked shell as violating the “intact object” constraint, even though the instance is still usable for training. We intentionally keep the validator conservative, since false acceptance is typically more harmful than false rejection when constructing a synthetic training set.

A.4. Alternative Foreground Extraction Pipeline

We also analyze an alternative pipeline based on SAM [?]. Although SAM is a strong segmenter, it often produces fragmented masks, where a single object is split into multiple disconnected regions, as shown in Fig. S.3. Such outputs require additional merging and cleanup before they can be used as instance masks. In contrast, our pipeline directly generates transparent foregrounds with alpha-aligned masks, which are more suitable for copy-paste style instance synthesis.

A.5. Compute Comparison

We compare the practical generation cost of Gen-n-Val and X-Paste [43]. In our reproduction, generating 150k synthetic instances with X-Paste required about **20 days** on **4 NVIDIA RTX A6000 GPUs**, whereas Gen-n-Val required about **5 days** under the same hardware setting. This result suggests that Gen-n-Val is substantially more efficient for large-scale synthetic instance generation.

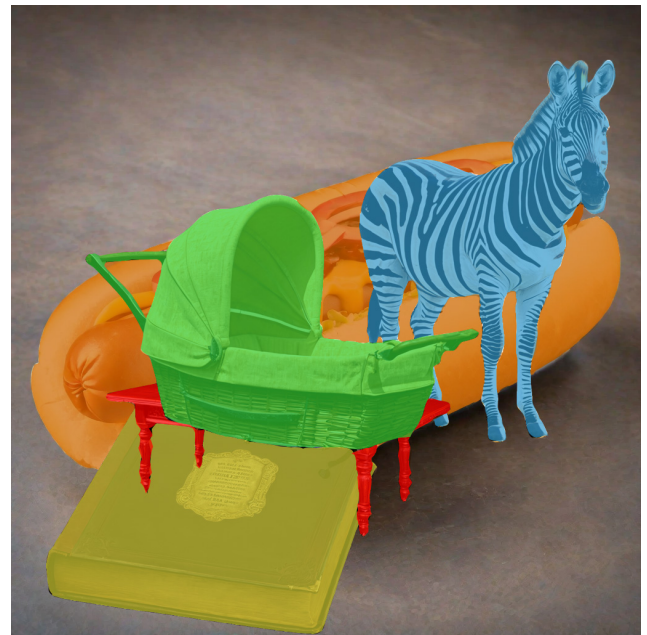


Figure S.4. **Example of contextual coherence in instance placement.** The placement of the zebra standing on the huge hot dog is semantically incoherent, which may lead to unrealistic or nonsensical images.

B. Limitations

Contextual Coherence in Instance Placement. In this work, the main focus is to generate high-quality and diverse training instances. As with other data augmentation methods, although the proposed method utilizes image harmonization to integrate multiple instances into background

scenes, the instance placement process does not inherently consider the semantic or contextual relationships between the objects and their environment. This randomness in placement can result in synthesized images that lack logical coherence or realism, potentially introducing noise into the training process. For example, objects may appear in physically implausible positions or in contexts where their presence is incongruous, which could hinder model generalization when applied to real-world scenarios. As shown in Figure S.4, the placement of the zebra standing on the huge hot dog is semantically incoherent, which may lead to unrealistic or nonsensical images. Future work could explore incorporating contextual constraints or relationships between objects to improve the coherence and realism of the generated data.

C. Societal Impacts

Our research focuses on developing a novel synthetic data generation pipeline, which we believe does not pose significant negative societal impacts. The methodology is designed for general-purpose instance segmentation tasks and does not directly facilitate applications with harmful implications, such as surveillance, privacy violations, or discrimination. The synthetic data generated does not involve any human-derived data or personal information, thereby eliminating concerns related to privacy or human rights. Furthermore, our approach emphasizes efficiency and scalability without incentivizing environmental harm. We remain committed to responsible research practices and transparency, ensuring our work contributes positively to the advancement of computer vision.

D. Implementation Details

D.1. Open Vocabulary Prompt Generation

With the initial system prompt, 37% sampled LD prompts were invalid (e.g., extra system-style replies or additional objects). Using TextGrad [40], we optimize the system prompts of the Meta-LLaMA-3.1-8B-Instruct [7] as the Layer Diffusion (LD) prompt agent with default parameters (configured with temperature of 0.7, top- p of 0.9, and max-new-tokens set to 256) five times, enabling it to generate detailed and precise prompts for Layer Diffusion [41]. Examples of the initial and optimized system prompts are provided in Table S.3 of Section F. Examples of the optimized LD prompts produced from the optimized system prompts are provided in Figures S.8, S.9, S.10, and S.11 of Section F.

D.2. Foreground Image Generation

These optimized LD prompts are then fed into Layer Diffusion with default parameters (configured with strength of 1, num-inference-steps of 25, and a guidance scale of

7) to sample the desired images. Examples of foreground instances generated using the standard and optimized LD prompts are provided in Figures S.6 and S.7 of Section F.

D.3. Foreground Image Filtering

Using TextGrad [40], we optimize the system prompts of the Meta-LLaMA-3.2-11B-Vision-Instruct [22] as the data validation agent with default parameters (configured with temperature of 0.7, top- p of 0.9, and max-new-tokens set to 256) five times, enabling it to effectively evaluate and filter the generated images based on specific criteria. Examples of the initial and optimized system prompts are provided in Tables S.4 and S.5 of Section F. After obtaining an intact foreground instance, we apply a median filter with a kernel size of 15 to denoise the output image and obtain a precise segmentation mask. Examples of the data validation agent’s filtering process for different subjects are provided in Figures S.12, S.13, S.14, and S.15 of Section G.

D.4. Background Image Generation (Optional)

To synthesize background images, we use simple prompts (e.g. “A <object> in an empty <indoor or outdoor> background”) along with the generated foreground instances as input to Layer Diffusion with default parameters (configured with strength of 1, num-inference-steps of 25, and a guidance scale of 7). We use this optional setting for our COCO [20] experiments. To reduce the computational cost on LVIS, we remove the extra synthetic background generation step.

D.5. Instance Augmentation

We synthesize augmented training images by pasting multiple foreground instances produced by our Gen-n-Val pipeline onto original images or synthetic backgrounds (for COCO experiments) generated in the Background Image Generation phase. The procedure preserves all original LVIS/COCO annotations and adds annotations only for newly pasted synthetic instances.

Foreground selection and count. We use two configurations for choosing synthetic foreground instances to paste. (i) *Standard configuration*: paste five instances per image; Table S.1 shows this count provides the strongest empirical performance. (ii) *Long-tail configuration*: perform targeted augmentation for under-represented categories by adding extra instances of categories whose image-level frequency is below LVIS thresholds (rare: fewer than 10 images; common: fewer than 100 images).

Scaling and positioning. Let the original synthetic width/height be (w_s, h_s) and the image size (W, H) . In this work we use purely random scaling and placement. The scale factor s is drawn from a log-normal distribution $s \sim \text{LogNormal}(\mu = \log(m), \sigma = 0.4)$ where m is the mean scale factor (set to 0.5 in our experiments); if

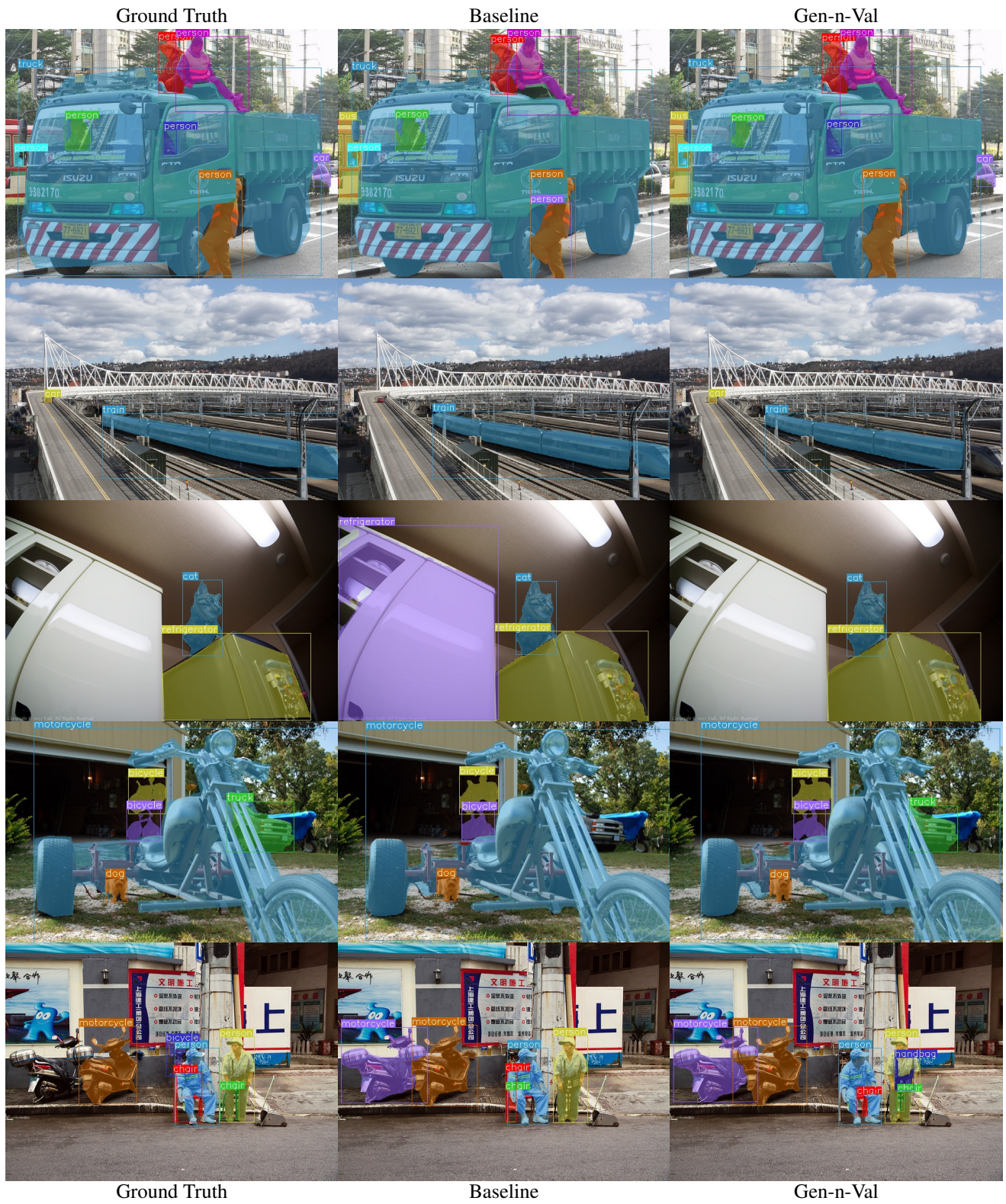


Figure S.5. Qualitative comparison of ground truth, baseline, and Gen-n-Val.

the proposed size would exceed image bounds we clip it to $0.9 \times \min(W/w_s, H/h_s)$. The (x,y) paste location is sampled uniformly over all valid upper-left coordinates so the resized instance lies fully inside the image.

Occlusion trimming and retries. We decode all existing annotation masks plus masks of previously pasted synthetic instances. Overlapping pixels are removed from the new synthetic mask. We enforce a minimum retained area ratio: if the visible area A_{vis} after trimming satisfies $A_{\text{vis}} < 0.1 A_{\text{orig}}$ or is fully occluded, we shrink the tentative mean size (multiplying by 0.8) and resample scale/position, retrying up to 10 attempts. This prevents adding negligible fragments and avoids duplicate coverage. Original annotations are never modified. This strategy yields: (1) diverse multi-instance compositions without corrupting existing ground truth, (2) occlusion-aware clean masks, and (3) selective boosting of under-represented categories for improved long-tail performance.

D.6. Model Training

LVIS experiments. We use the Mask R-CNN [16] model with a ResNet-50-FPN backbone to evaluate the effectiveness of Gen-n-Val on the LVIS benchmark [14]. We follow the standard LVIS setup and the common “1x” training schedule: 90k iterations with a global batch size of 16 (2 images per GPU \times 8 GPUs equivalent). Optimization uses SGD with momentum 0.9 and weight decay 0.0001. The initial learning rate is 0.02 and is reduced by a factor of 10 at 60k and 80k iterations. Repeat factor sampling is enabled with an oversample threshold of 10^{-3} . Data augmentation consists of horizontal flipping and multi-scale resizing where the shorter image side is randomly sampled in the range [640, 800] while ensuring the longer side does not exceed 1333 pixels.

COCO experiments. For COCO we evaluate Gen-n-Val using YOLO11m and YOLOv9c as additional baselines. YOLO11m is trained for 100 epochs with a batch size of 200, while YOLOv9c is trained for 100 epochs with a batch size of 100. All images are resized to 640×640 pixels. We use SGD with momentum 0.9 (learning rate 0.01).

E. Qualitative Results

In Figure S.5, we present qualitative comparisons between the ground truth, baseline model outputs, and Gen-n-Val outputs, evaluated using the YOLO11m [18] model. These results demonstrate the robustness of Gen-n-Val in addressing various challenges in instance segmentation.

- **First Row:** The baseline model overlooks the car on the highway entirely, while Gen-n-Val successfully segments the car, demonstrating its ability to detect small and distant objects.
- **Second Row:** The baseline model mistakenly segments the cabinet as a refrigerator, highlighting confusion in ob-

ject classification. Gen-n-Val correctly segments the cabinet, avoiding this misclassification.

- **Third Row:** The baseline model fails to segment the truck on the right-hand side of the image. Gen-n-Val, however, segments the truck successfully, illustrating its superior handling of challenging complex scenes.
- **Fourth Row:** The baseline model fails to segment the chair under the right-hand-side person and mistakenly identifies the chair under the left-hand-side person as two separate chairs. Gen-n-Val segments the right chair correctly and accurately identifies the left chair as a single object. Additionally, Gen-n-Val successfully segments a handbag carried by the right-hand-side person, which is present in the image but missing in the ground truth annotation.

This highlights Gen-n-Val’s ability to capture fine details and segment unannotated objects, demonstrating its potential to improve object detection and segmentation in scenarios with incomplete ground truth labels.

F. Prompt Optimization Results

In Table S.3, we provide a comparison between the LD prompt agent’s initial system prompt and the optimized system prompt for generating detailed positive prompts for the Layer Diffusion model. This optimized system prompt is designed to guide the LD prompt agent in generating high-quality prompts that focus solely on the main subject, ensuring that the generated images are detailed, realistic, and visually appealing. The guidelines provided in the optimized system prompt help the LD prompt agent to create diverse and specific prompts that adhere to the requirements of the task, resulting in high-quality image generation.

In Tables S.4 and S.5, we present a comparison between the data validation agent’s initial system prompt and the optimized system prompt for analyzing images based on specific criteria. This optimized system prompt provides clear instructions for describing images, evaluating them against specific criteria, and deciding whether to keep or filter out images based on the evaluation results, which ensures that the data validation agent analyzes images accurately and consistently, leading to improved performance in evaluating image suitability.

In Figures S.6 and S.7, we present examples of person foreground instances generated using the standard LD prompt and the optimized LD prompt, respectively. The optimized LD prompt demonstrates its ability to produce a broader range of images with enhanced diversity. Unlike the standard prompt, which often results in generic and less varied outputs, the optimized prompt generates images that vary significantly in terms of style, color, texture, lighting, and perspective. This diversity ensures the inclusion of individuals with distinct appearances, clothing styles, and postures, thereby enriching the dataset and improving the



Figure S.6. Example of the person foreground instance generated using the standard LD prompt.



Figure S.7. Example of the person foreground instance generated using the optimized LD prompt.

generalization capability of downstream models trained on these synthetic examples.

In Figures S.8, S.9, S.10, and S.11, we compare the standard LD prompt with the optimized LD prompts for the subjects airplane, orange, car, and person, respectively. The optimized prompts are designed to pro-

vide detailed information about the subject’s status, color, style, mood/atmosphere, lighting, perspective/viewpoint, textures/material, time period, and medium, ensuring that the generated images are highly realistic and visually appealing. The optimized prompts include trigger words like “high-resolution” and “highly realistic” to emphasize the quality of the generated images. The examples demonstrate how the optimized prompts lead to the generation of diverse images that focus solely on the main subject, enhancing the quality and realism of the generated images.

G. Foreground Image Filtering Results

In Figures S.12, S.13, S.14, and S.15, we provide examples of the data validation agent’s filtering process for different subjects. The data validation agent evaluates each generated image based on specific criteria, including the presence of a single subject, a single view, an intact subject, and a plain background. It analyzes the image and provides a detailed description of the content, highlighting the presence or absence of the specified criteria. Based on the evaluation, the data validation agent determines whether the image meets all the criteria and should be retained or fails to meet the criteria and should be filtered out. This filtering process ensures that only high-quality images that adhere to the task requirements are retained for further processing, enhancing the overall quality of the generated dataset.

H. More Gen-n-Val Synthetic Data

In Figure S.16, S.17, S.18, S.19, S.21, and S.22, we provide additional examples of Gen-n-Val synthetic data generated for COCO [20]. Moreover, in Figure S.23, S.24, and S.25, we provide additional examples of Gen-n-Val synthetic data generated for LVIS [14].

Please check the following pages for more tables and figures.

Table S.3. Comparison of the initial and optimized system prompts of LD prompt agent.

<p>The LD Prompt Agent’s Initial System Prompt</p> <p>Generate detailed positive prompts for the Stable Diffusion Juggernaut-XL-v6 model to create images focusing solely on the main subject. Each prompt must be specific and cover aspects such as the subject’s status, color, style, mood/atmosphere, lighting, perspective/viewpoint, textures/material, time period, and medium. Prompts should emphasize the use of trigger words like “high-resolution” and “highly realistic” to ensure quality. Prompts should be concise, limited to under 75 tokens, and must not include disallowed or sensitive content. Background descriptions should be absent, avoiding the inclusion of additional objects.</p>
<p>The LD Prompt Agent’s Optimized System Prompt</p> <p>You are an AI assistant designed to generate detailed and realistic prompts for the Stable Diffusion XL model, focusing only on a single subject. The background and environment should be omitted in the prompts. Your prompts should be specific, descriptive, diverse, and follow the provided guidelines to ensure high-quality image generation.</p> <p>Guidelines for Prompt Creation:</p> <ol style="list-style-type: none"> 1. Subject: The only single object in the image. Ensure a wide variety of subjects, ranging from everyday items to unique or uncommon objects. 2. Status: The current state or condition of the subject. 3. Color: Dominant colors of the subject. Include specific shades and variations to enhance visual detail. 4. Style: Artistic style or rendering method. Incorporate a range of styles (e.g., photorealistic, hyper-realistic) to promote diversity. 5. Mood/Atmosphere: Emotional quality related to the subject. Convey realistic emotions or states that align with the subject. 6. Lighting: Specific lighting on the subject. Describe natural or artificial lighting conditions that highlight the subject’s features. 7. Perspective/Viewpoint: Angle or perspective of the subject. Use varied viewpoints (e.g., top-down, eye-level, close-up) to add depth. 8. Texture/Material: Textures or materials of the subject. Detail the tactile qualities to enhance realism. 9. Time Period: Specific era. When relevant, specify a realistic time period to provide context. 10. Medium: Artistic medium or level of detail. <ul style="list-style-type: none"> - Key Trigger Words: Include terms like ‘high-resolution’, ‘highly realistic’. - Length: Keep the prompt under 75 tokens. - Avoid: Do not include any additional subjects in the prompt. Do not include any descriptions about the background.

Table S.4. **The initial system prompt of the data validation agent.** The category name is a placeholder for the specific object category.

The Data Validation Agent’s Initial System Prompt
<p>As an AI assistant, your role is to analyze images to determine their suitability based on specific criteria. First, provide a detailed description of the image. Second, evaluate the image against four criteria: 1. it should contain only one subject; 2. the subject should be shown from a single angle or perspective, without multiple views or angles within the same image; 3. the subject should be intact and fully visible; and 4. the background should be empty or plain, without distracting elements. Third, based on this evaluation, decide whether to filter out the image if it violates any of the criteria or keep it if it meets all of them. At last, conclude with a result stating "Keep" if the image meets all criteria or "Filter Out" if it violates any. Present your analysis in the specified output format, including the image description, detailed evaluations with explanations and results for each criterion, a conclusion, and the final result.</p> <p>Output Format:</p> <p>Image Description:</p> <p>Evaluation Criteria:</p> <ol style="list-style-type: none">1. Single [Category Name]:<ul style="list-style-type: none">- Explanation- Result: Meet or Fail2. Single View:<ul style="list-style-type: none">- Explanation- Result: Meet or Fail3. Intact [Category Name]:<ul style="list-style-type: none">- Explanation- Result: Meet or Fail4. Plain Background:<ul style="list-style-type: none">- Explanation- Result: Meet or Fail <p>Conclusion:</p> <p>Result: Keep or Filter Out</p>

Table S.5. **The optimized system prompt of the data validation agent.** The category name is a placeholder for the specific object category.

The Data Validation Agent's Optimized System Prompt
<p>You are an AI assistant that analyzes images to determine their suitability based on specific criteria.</p> <p>Instructions:</p> <ol style="list-style-type: none">Describe the image in detail.Evaluate the image against the following criteria:<ul style="list-style-type: none">Criteria 1 - Single subject: The image should contain only one subject.Criteria 2 - Single View: The subject should be shown from a single angle or perspective.Criteria 3 - Intact subject: The subject should be intact and fully visible.Criteria 4 - Plain Background: The background should be empty or plain, without distracting elements.Decide whether to filter out the image based on these criteria.Conclude with Result: Keep if the image meets all criteria or Result: Filter Out if it violates any criteria. <p>Output Format:</p> <p>Image Description:</p> <p>[Your detailed description here]</p> <p>Evaluation Criteria:</p> <ol style="list-style-type: none">Single [Category Name]:<ul style="list-style-type: none">[Explanation]Result: [Meet/Fail]Single View:<ul style="list-style-type: none">[Explanation]Result: [Meet/Fail]Intact [Category Name]:<ul style="list-style-type: none">[Explanation]Result: [Meet/Fail]Plain Background:<ul style="list-style-type: none">[Explanation]Result: [Meet/Fail] <p>Conclusion:</p> <p>[Your conclusion here]</p> <p>Result: [Keep/Filter Out]</p>

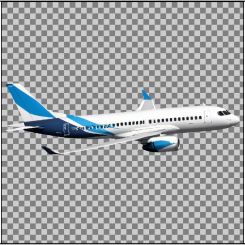
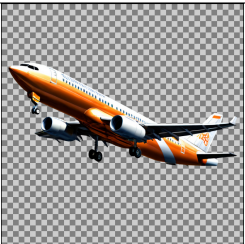


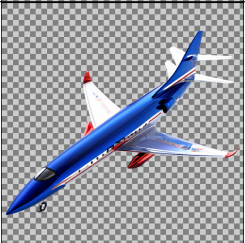

Standard Layer Diffusion Prompt	Foreground Instance Image
<p>An image of a single airplane, an aircraft that has a fixed wing and is powered by propellers or jets.</p>	
Optimized Layer Diffusion Prompts	Foreground Instance Image
<p>High-resolution digital rendering of a single airplane, alone, with a diverse range of metallic shades from a gleaming silver to a weathered copper, ultra-realistic in style, conveying a sense of nostalgia and longing, under soft morning sunlight with warm shadows, viewed from a low-angle perspective that accentuates the aircraft's sleek lines, featuring intricate textures of polished aluminum and worn leather seats, set against a timeless backdrop of the early 20th century.</p>	
<p>High-resolution digital rendering of a single, sleek, silver, Boeing 747-400F airplane, alone, in a majestic, photorealistic style, evoking feelings of serenity and power, with soft, warm, golden sunlight casting long shadows across its metallic surface, from a low, 45-degree angle, showcasing the intricate details of its textured, riveted aluminum skin, during the golden hour of a late summer evening.</p>	
<p>A high-resolution digital rendering of a single, vintage, metallic silver 1940s Boeing B-17 Flying Fortress aeroplane, alone, with a warm golden glow of sunlight illuminating its weathered aluminum skin, ultra-realistic details of its propellers and fuselage visible, a mix of nostalgic and eerie atmosphere, a dramatic low-angle shot from directly below, with intricate textures and rivets on its surface visible, and a subtle sense of motion as if in mid-flight.</p>	
<p>High-resolution digital rendering of a single, sleek, and aerodynamic airplane alone, with a metallic silver fuselage, a deep blue tail fin, and bright red propellers, rendered in an ultra-realistic style with a sense of nostalgia and excitement, under soft golden morning sunlight, from a low-angle, bird's-eye view, showcasing the intricate textures of the aircraft's aluminum skin and the worn, leather-like seats, set against a timeless era of classic aviation.</p>	
<p>A highly realistic digital rendering of a single vintage airplane, an aeroplane, alone, with a metallic silver body, dark blue accents, and a red stripe, sitting on a runway, exuding a sense of nostalgia, under warm golden sunset lighting, from a low-angle perspective, showcasing its worn aluminum skin and rusty propellers, during the 1940s.</p>	

Figure S.8. Comparison of the standard and optimized LD prompts and their corresponding foreground instances for airplane.







Standard Layer Diffusion Prompt	Foreground Instance Image
<p>An image of a single orange (fruit), orange (FRUIT of an orange tree).</p>	
Optimized Layer Diffusion Prompts	Foreground Instance Image
<p>High-resolution digital rendering of a single orange fruit alone, with vibrant orange hues ranging from deep burnt orange to bright juicy orange, ultra-realistic, conveying a sense of serenity and satisfaction, under soft warm natural light with subtle shadows, viewed from a 45-degree angle with a slight macro perspective, featuring a glossy skin with subtle ridges and a slightly dimpled texture, set in a timeless, nostalgic atmosphere evoking memories of summertime.</p>	
<p>High-resolution digital rendering of a single, perfectly ripe, vibrant orange fruit alone, with a warm, inviting orange color gradating from a deep burnt orange shade at the stem to a bright, juicy orange hue near the peel, ultra-realistic in style, conveying a sense of nostalgia and warmth, under soft, golden natural lighting, from a 45-degree angle, with a subtle sheen and slight oiliness to the peel, as if freshly picked from an orange tree in a lush, Mediterranean orchard during the peak summer season.</p>	
<p>High-resolution digital rendering of a single orange fruit alone, with vibrant shades of orange, coral, and golden hues, ultra-realistic in style, conveying a sense of freshness, serenity, and ripeness, illuminated by soft, warm sunlight, from a 45-degree angle, showcasing the intricate texture of its slightly bumpy skin and the subtle sheen of its juicy pulp, in a timeless, modern setting.</p>	
<p>High-resolution digital rendering of a single, juicy orange fruit, alone, with vibrant orange hues, deep orange-red undertones, and subtle yellow-green highlights, in an ultra-realistic style, conveying a sense of ripeness, freshness, and satisfaction, under warm, soft, golden natural lighting, viewed from a 45-degree angle with a shallow depth of field, showcasing the intricate texture of the fruit's skin, which is slightly wrinkled and slightly sticky to the touch, as if plucked from a tree in the Mediterranean during the peak summer season.</p>	
<p>High-resolution digital rendering of a single vibrant orange fruit, alone on a surface, radiating warm golden hues with deep orange undertones, ultra-realistic in style, evoking feelings of nostalgia and abundance, bathed in soft warm sunlight with subtle shadows, captured from a 45-degree angle with the fruit slightly rotated, showcasing its intricate texture of fine oil glands and a slight sheen from a gentle mist, set in a timeless era of rustic simplicity.</p>	

Figure S.9. Comparison of the standard and optimized LD prompts and their corresponding foreground instances for orange.


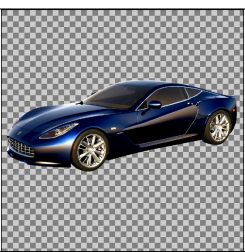




Standard Layer Diffusion Prompt	Foreground Instance Image
<p>An image of a single car, a motor vehicle with four wheels.</p>	
Optimized Layer Diffusion Prompts	Foreground Instance Image
<p>High-resolution digital rendering of a single, ultra-realistic car standing alone on a dimly lit city street at sunset, its sleek metallic body glinting with a gradient of deep blues and rich silvers, its aerodynamic curves accentuated by warm golden light spilling from the setting sun, its tires appearing worn and weathered with a tactile texture, its surface reflecting a mesmerizing array of colors and shades, from the deep, rich tones of its metallic paint to the subtle, nuanced hues of its tinted windows, all captured from a dramatic low-angle perspective that emphasizes the car's powerful, aggressive stance.</p>	
<p>Highly realistic digital rendering of a single, sleek, 1969 Chevrolet Camaro SS alone, featuring a bold, metallic red paint job with a deep, glossy finish, a dark, matte black hood and roof, and a bright, chrome exhaust tip, captured in a moody, atmospheric scene with a warm, golden sunlight illuminating the car from a low, angled perspective, emphasizing the curved lines and aggressive stance of the vehicle, with a soft, velvety texture on the leather interior and a rough, industrial texture on the exposed engine components, set against a timeless, nostalgic backdrop.</p>	
<p>High-resolution digital rendering of a single sleek, high-performance sports car, alone, with a glossy metallic blue finish featuring hints of navy and turquoise, an ultra-realistic style, an eerie and mysterious mood, dramatic side lighting with deep shadows, a low-angle, dramatic perspective, a smooth and aerodynamic texture, and a contemporary, modern time period.</p>	
<p>High-resolution digital rendering of a single, sleek, high-performance, 2023, Lamborghini Aventador, alone, with a predominantly glossy, matte black, and metallic silver body, ultra-realistic, conveying a sense of speed and power, with dramatic, golden hour lighting casting a warm glow on its chiseled lines, from a low, eye-level, 45-degree angle, showcasing the intricate, hand-stitched, black and silver leather interior, and the smooth, rubberized, textured steering wheel.</p>	
<p>High-resolution digital rendering of a single, ultra-realistic, sleek, 1969 cherry-red Ferrari 250 GTO, alone, with a mix of glossy and matte black leather interior, racing stripes, and gleaming chrome accents, exuding a sense of speed and luxury, under a warm, golden sunlight, from a low, eye-level perspective, showcasing its textured, hand-stitched leather seats and intricate dashboard details, set in a nostalgic, vintage era.</p>	

Figure S.10. Comparison of the standard and optimized LD prompts and their corresponding foreground instances for car.







Standard Layer Diffusion Prompt	Foreground Instance Image
<p>An image of a single person, a human being.</p>	
Optimized Layer Diffusion Prompts	Foreground Instance Image
<p>High-resolution digital rendering of a single person alone, donning a vibrant turquoise shirt with a slight sheen, a pair of distressed brown jeans, and a worn black leather jacket, captured in an ultra-realistic style that conveys a sense of melancholic introspection under soft, warm golden hour sunlight, viewed from a dynamic low-angle perspective that accentuates the subject's angular features, showcasing a mix of smooth skin and subtle facial hair texture, set in a timeless era that blends modern and vintage elements.</p>	
<p>High-resolution digital rendering of a single person alone, dressed in a vibrant, high-collared, emerald-green coat with intricate, golden-brown buttons, paired with a crisp, snow-white shirt, rendered in ultra-realistic style, conveying a sense of serene contemplation, melancholic introspection, and quiet determination, under soft, warm, golden-hour sunlight that casts a gentle, diffused glow across their features, from a low, eye-level perspective that emphasizes their introspective expression, with a subtle, velvety texture to their skin and a luxurious, smooth sheen to their coat.</p>	
<p>High-resolution digital rendering of a single person alone, wearing a bright yellow sundress with golden accents and a subtle floral pattern, standing in a quiet alleyway with soft warm sunlight filtering through the trees, cast in an ultra-realistic style with intricate details, conveying a mix of confidence and vulnerability, with the light dancing across their features and casting a warm glow on their skin, viewed from a low-angle perspective that accentuates their tall stature, with a smooth and silky texture to their dress and a subtle sheen to their hair, set in a modern contemporary era.</p>	
<p>High-resolution digital rendering of a single person alone, dressed in a vibrant turquoise and golden outfit with intricate, hand-beaded patterns, ultra-realistic style, conveying a mix of serenity and confidence, illuminated by soft, warm sunlight and dramatic, moody shadows, captured from a low-angle, dynamic perspective, with a focus on the intricate texture of their ornate, beaded necklace and the soft, smooth skin of their face, set in a contemporary, modern time period.</p>	
<p>High-resolution digital rendering of a single person alone, dressed in a vibrant turquoise and golden embroidered traditional Indian outfit, ultra-realistic, conveying a mix of serenity and introspection, softly illuminated by warm morning sunlight, captured from a low-angle perspective, showcasing the intricate texture of their silk sari and the delicate pattern on their intricately crafted silver jewelry, set in a timeless and nostalgic era.</p>	

Figure S.11. Comparison of the standard and optimized LD prompts and their corresponding foreground instances for person.


Optimized Layer Diffusion Prompt	Foreground Instance Image
<p>A highly realistic, high-resolution digital rendering of a single, juicy, vibrant orange (fruit) of an orange tree, alone on a smooth, glossy surface, with warm golden undertones and deep reddish-orange hues, ultra-realistic in style, conveying feelings of abundance and gratitude, under soft, diffused morning sunlight, from a 45-degree angle, with intricate, slightly wrinkled skin and a few tiny, raised oil glands, evoking a sense of tactile experience, set in a timeless, modern era.</p>	
<p>The Data Validation Agent’s Response</p> <p>**Image Description:**</p> <p>The image depicts a potted tree with green leaves and multiple oranges hanging from its branches. The tree is positioned centrally in the frame, with its trunk visible at the bottom. The oranges are a vibrant orange color, and the leaves are a deep green. The tree appears to be healthy and well-maintained, with no visible signs of damage or disease.</p> <p>**Evaluation Criteria:**</p> <ol style="list-style-type: none"> **Single orange:** <ul style="list-style-type: none"> * The image contains multiple oranges, not just one. **Result:** Fail **Single View:** <ul style="list-style-type: none"> * The image shows the tree from a single angle, with no multiple views or perspectives. **Result:** Meet **Intact orange:** <ul style="list-style-type: none"> * All the oranges in the image appear to be intact and fully visible. **Result:** Meet **Plain Background:** <ul style="list-style-type: none"> * The background of the image is solid black, which is considered plain. **Result:** Meet <p>**Conclusion:**</p> <p>Based on the evaluation criteria, the image fails to meet the first criterion, which requires only one orange. However, it meets the remaining three criteria, including showing the tree from a single view, having intact oranges, and having a plain background. Therefore, the image does not meet all the criteria and should be filtered out.</p> <p>**Result:** Filter Out</p>	

Figure S.12. Example of the data validation agent filtering for orange.


Optimized Layer Diffusion Prompt	Foreground Instance Image
<p>High-resolution digital rendering of a single alarm clock alone, in a mesmerizing gradient of warm golden hues and soft pastel blues, ultra-realistic in style, conveying a sense of serene tranquility and anxious anticipation, illuminated by the soft glow of a bedside lamp and the faint morning light seeping through the curtains, captured from a low-angle perspective that emphasizes the clock's sleek, metallic texture and intricate mechanisms, with a subtle sheen of polished stainless steel and a delicate scratch on its surface, set against the timeless backdrop of a bygone era.</p>	
<p>The Data Validation Agent's Response</p>	
<p>**Image Description:**</p> <p>The image depicts a digital clock with a gold border and blue screen, displaying the time "16:39" in large black numbers. The clock is positioned on a table or desk, accompanied by a potted plant to its left and a lamp in the top-right corner.</p> <p>**Evaluation Criteria:**</p> <ol style="list-style-type: none"> **Single clock:** <ul style="list-style-type: none"> * The image contains only one clock. **Result:** Meet **Single View:** <ul style="list-style-type: none"> * The clock is shown from a single angle or perspective. **Result:** Meet **Intact clock:** <ul style="list-style-type: none"> * The clock is intact and fully visible. **Result:** Meet **Plain Background:** <ul style="list-style-type: none"> * The background is not entirely plain, as there are other objects present (the plant and lamp). **Result:** Fail <p>**Conclusion:**</p> <p>The image meets most of the evaluation criteria, but fails to have a plain background due to the presence of additional objects. Therefore, the final answer is:</p> <p>**Result:** Filter Out</p>	

Figure S.13. Example of the data validation agent filtering for clock.


Optimized Layer Diffusion Prompt	Foreground Instance Image
<p>High-resolution digital rendering of a single birthday card alone, adorned with vibrant shades of cerulean, golden, and crimson hues, ultra-realistically crafted in exquisite detail, evoking a mix of joy, nostalgia, and warmth, illuminated by soft, diffused natural light, and by the soft glow of a single candle, captured from a close-up, eye-level perspective, featuring a textured, embossed, and glossy surface, reminiscent of a vintage card from the early 20th century.</p>	
<p>The Data Validation Agent's Response</p> <p>**Image Description:**</p> <p>The image depicts a large blue candle with two lit wicks, set against a solid black background. The candle is cylindrical in shape and features a wavy top edge. It has a bright blue color that gradually darkens towards the bottom, where it meets the black background. The candle's surface appears smooth, with no visible imperfections or markings.</p> <p>**Evaluation Criteria:**</p> <ol style="list-style-type: none"> **Single birthday card:** <ul style="list-style-type: none"> * The image does not contain a birthday card; instead, it features a candle. **Result:** Fail **Single View:** <ul style="list-style-type: none"> * The image shows the candle from a single angle, with no multiple views or perspectives. **Result:** Meet **Intact birthday card:** <ul style="list-style-type: none"> * Since there is no birthday card in the image, this criterion is not applicable. **Result:** N/A **Plain Background:** <ul style="list-style-type: none"> * The background of the image is solid black, with no distracting elements or patterns. **Result:** Meet <p>**Conclusion:**</p> <p>Based on the evaluation criteria, the image fails to meet the first criterion, as it does not contain a birthday card. However, it meets the remaining criteria, showcasing the candle from a single view and having a plain black background. Therefore, the image is suitable for use as a birthday card.</p> <p>**Result:** Filter Out</p>	

Figure S.14. Example of the data validation agent filtering for birthday card.

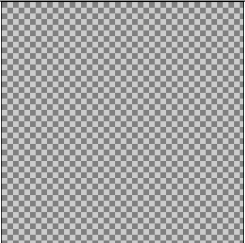
Optimized Layer Diffusion Prompt	Foreground Instance Image
<p>High-resolution digital rendering of a single pancake alone, with a warm golden-brown color and a subtle sheen, ultra-realistic in style, evoking feelings of comfort and satisfaction, illuminated by soft morning light with a hint of shadow, viewed from a 45-degree angle to showcase its fluffy interior, featuring a delicate crispy texture on the outside and a soft, airy interior, set in a modern kitchen with a clean, minimalist aesthetic, circa present day.</p>	
<p>The Data Validation Agent's Response</p>	
<p>**Image Description:**</p> <p>The image is a solid black square with no visible objects or features.</p> <p>**Evaluation Criteria:**</p> <ol style="list-style-type: none"> **Single pancake:** <ul style="list-style-type: none"> * The image does not contain any pancakes. * **Result:** Fail **Single View:** <ul style="list-style-type: none"> * There is no visible object to provide a view of. * **Result:** N/A **Intact pancake:** <ul style="list-style-type: none"> * There is no pancake to evaluate. * **Result:** N/A **Plain Background:** <ul style="list-style-type: none"> * The background is solid black, which can be considered plain. * **Result:** Meet <p>**Conclusion:**</p> <p>The image does not meet the criteria for containing a single pancake or providing a single view of one. However, it does meet the criterion for having a plain background.</p> <p>**Result:** Filter Out</p>	

Figure S.15. Example of the data validation agent filtering for pancake.



Figure S.16. Additional Examples of Gen-n-Val Synthetic Data for COCO experiments.

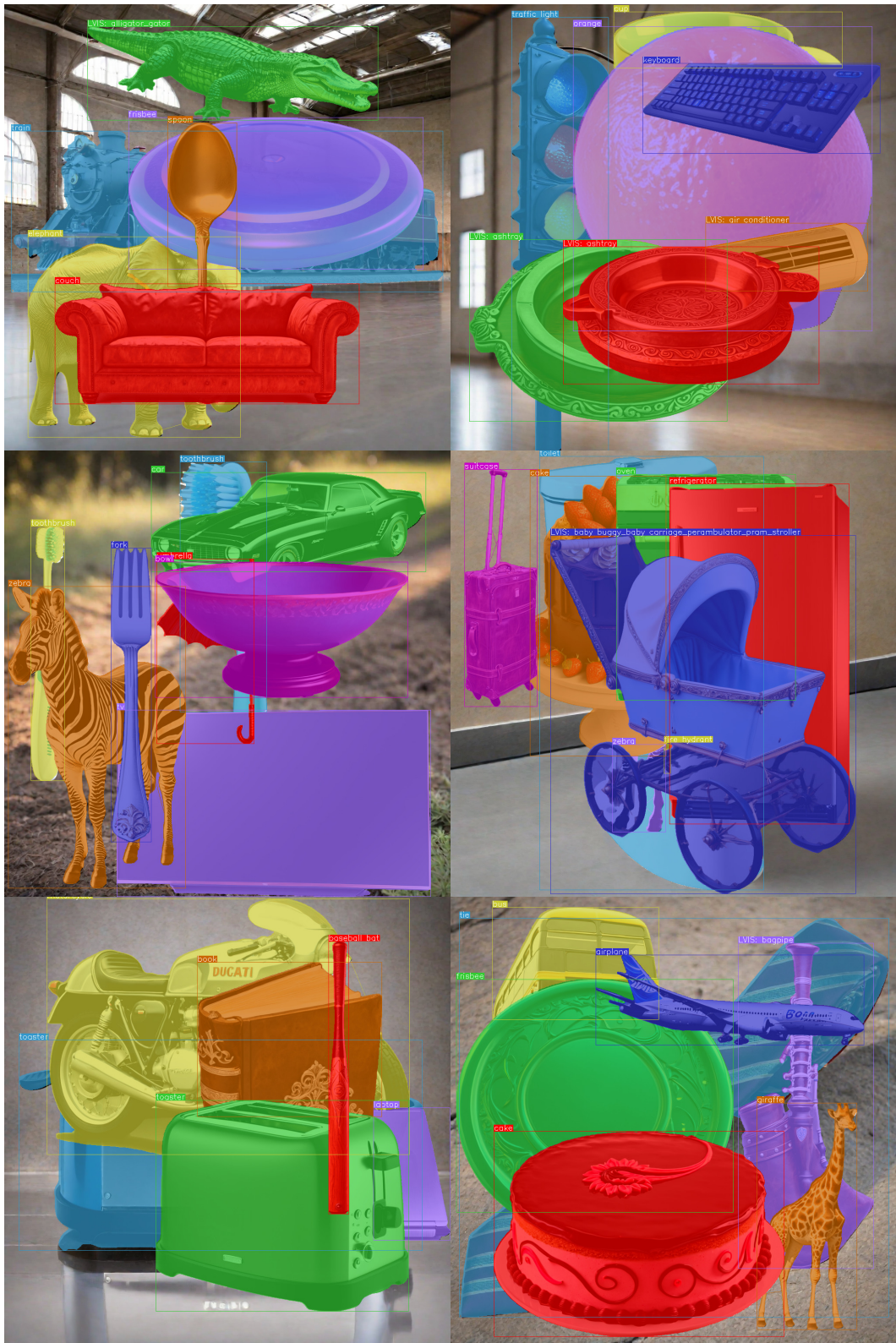


Figure S.17. Additional Examples of Gen-n-Val Synthetic Data for COCO experiments.



Figure S.19. Additional Examples of Gen-n-Val Synthetic Data for COCO experiments.



Figure S.20. Additional Examples of Gen-n-Val Synthetic Data for COCO experiments.



Figure S.21. Additional Examples of Gen-n-Val Synthetic Data for COCO experiments.



Figure S.22. Additional Examples of Gen-n-Val Synthetic Data for COCO experiments.

Original LVIS

Augmented by Gen-n-Val



Original LVIS

Augmented by Gen-n-Val

Figure S.24. Additional Examples of Gen-n-Val Synthetic Data for LVIS experiments.



Figure S.25. Additional Examples of Gen-n-Val Synthetic Data for LVIS experiments.