# UNSPOKEN, CONTROLLABLE, UNAVOIDABLE: IMPLICIT CO-OCCURRING OBJECTS IS A FUNDAMENTAL PROPERTY IN TEXT-TO-IMAGE GENERATION

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

Text-to-Image (T2I) models often generate objects that are not in the prompt: e.g., a cup on a table when the prompt is simply "a photo of a cup". We call those the Implicit Co-occurring Objects (ICO). In this work, we present the first systematic study of ICO across seven architecturally distinct T2I models spanning the years 2022-2026 with a strategic plan for the understanding and the control of them. We find that 34-50% of the generated images featuring common objects (56,000 across 80 categories), regardless of the model, contain at least one unprompted object, with strikingly consistent co-occurring patterns shared across the models. To understand where they come from and how models generate them, we investigate four hypotheses about the training data, text encoders, latent space trajectory planning, and sampling dynamics. Training data co-occurrence statistics are found to be the dominant factor, and the text encoder has a secondary effect. And the sampling parameters modulate the magnitude. Guided by the insights, we explore a group of controls of ICO and discover a simple yet effective two-component control method combining isolation phrasing with embedding steering that reduces the leakage rate to 0.19% with a 94.2% fidelity. For a theoretical understanding, we show that having ICO is a fundamental property of the training data, labels, and the objective of T2I models: with co-occurrence statistics existing in the natural image distribution, there is a trade-off between the model-vs-data distribution divergence and the rate of ICO, meaning we cannot freely suppress ICOs without a loss of generation fidelity. Finally, we show that ICO carries a downstream cost: object detectors trained on T2I-generated images exhibit up to an 18.8% gap in recall compared to ICO-free data. We are the first to establish a comprehensive understanding of ICO in T2I models. Our findings provide new insight into this intrinsic property of diffusion-based image synthesis.

## 1 INTRODUCTION

**The phenomenon.** Text-to-Image (T2I) models (Yang et al., 2025; Cao et al., 2025) generate high-quality images based on a user's prompt: for example, a photorealistic image of a coffee cup. But a careful user would notice often times the generated images contain additional, unrequested objects: a table where the coffee cup sits or a computer monitor by the side of the coffee cup (Figure 1, note they are also full of coffee).

Prior work has studied unexpected behavior of generative models, but this is something different. This is not *bias* (Afreen et al., 2025), e.g., generating a certain type of coffee cup with gender or demographics. This is not *hallucination* (Huang et al., 2025a), e.g., generating contradicting or wrong objects (a wine glass instead). Oftentimes this phenomenon goes under the radar, as the user may not think this generation as "wrong" or "undesirable", yet the table and the monitor are never requested. We call these the Implicit Co-occurring Objects (ICOs) in T2I generations.

**Why it matters?** Although looking harmless, we argue that this phenomenon can have a significant impact for tasks that leverage generated images. Recently, T2I models have become popular for generating synthetic images as training data for downstream perception tasks (He et al., 2023). A model that is trained on such data will silently pair prompted objects with those unspoken objects.



Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.




Figure 1: All T2I models we tested generate implicit co-occurring objects: the above prompt is for "a photo of cup". This happens for all common objects (last column for prompt-vs-objects co-occurrence). Clean generation should produce a clean diagonal co-occurrence matrix.

Critically, ICOs cannot be easily dismissed with better generation models and mechanisms: 1) More details in prompts can reduce but never eliminate unprompted objects. 2) We found this phenomenon empirically across 7 widely adopted, architecturally distinct, and size-variant models from 2022 to 2026, with no clear correlation with model architecture, hyperparameters, size, or recency.

**In this paper,** we conduct a systematic study on the ubiquitous yet under-studied ICOs:

- We characterize the phenomenon across seven architecturally distinct models using four independent probes and customized metrics (Sec. 3).
- We investigate the source and mechanisms of generating ICOs with four non-exclusive hypotheses spanning training data, text encoder, sampling dynamics, and measure contributions (Sec. 4).
- We control the leakage using mechanisms derived from Sec. 4 that achieve competitive results with SOTA concept-erasure methods (Sec 5).
- We prove that, in theory, any inference-time method trying to reduce the leakage below the real-world co-occurrence rate will drift away from the real photo distribution by a margin resulting in fidelity loss, which is confirmed by our tested method empirically. This is a fundamental property of text-to-image generation (Sec 6).
- We study the impact of having the leakage by showing substantial performance gaps in object detectors trained with ICOs vs clean data (Sec 7).

To the best of our knowledge, this is the first work to establish a comprehensive understanding of ICOs in T2I generation. Our findings provide new tools, theory, and insights into this quantifiable and intrinsic property for diffusion-based image synthesis work.

## 2 Related Work

**Text-to-Image diffusion models and their characteristics**. We focus our study on T2I generation by diffusion and flow-matching models (Ho et al., 2020; Lipman et al., 2023; Liu et al., 2023): 1) CLIP-conditioned (Radford et al., 2021) latent U-Nets: SD1.5 (Rombach et al., 2022), SDXL (Podell et al., 2024); 2) T5-conditoned (Raffel et al., 2020) diffusion transformers (Peebles & Xie, 2023; Esser et al., 2024): PixArt-$\alpha$ (Chen et al., 2024) and Flux-dev (Black Forest Labs, 2024); 3) Multimodal-LLM-conditioned flow transformers : Qwen-Image (Wu et al., 2025), Z-Image-Turbo (Z-Image Team, 2025), Krea 2 (Lee et al., 2026). We use their 8-step distilled versions. The T2I models has various LLM text encoders (Saharia et al., 2022), dense synthetic captions (Betker et al., 2023), distillation (Meng et al., 2023), and few-step sampling.

**Bias and hallucination in vision generative models**. Generative models reproduce demographic stereotypes (Bianchi et al., 2023; Luccioni et al., 2023; Cho et al., 2023; Nadeem et al., 2025), due to biased data (Zhao et al., 2017; Seshadri et al., 2024). Wu et al. (2024b) attribute gender-dependent objects to text embeddings. Hallucination can take different forms (Aithal et al., 2024; Lim et al., 2025; Bansal et al., 2025). VLMs output co-occurring but absent objects in hallucinated answers (Rohrbach et al., 2018; Li et al., 2023; Zhou et al., 2024). LLMs also get hallucination and bias from data (Huang et al., 2025a; Caliskan et al., 2017). Theory shows that calibrated generators must hallucinate (Kalai & Vempala, 2024; Kalai et al., 2026). We make the T2I analogue: ICOs are unavoidable in a well-trained T2I model.

**Object co-occurrence and context**. Context is key to recognition (Biederman et al., 1982; Oliva & Torralba, 2007; Torralba, 2003; Rabinovich et al., 2007; Divvala et al., 2009; Hu et al., 2018).

Table 1: Text-to-Image models studied in this paper: they are highly popular and widely adopted.

| Model | Architecture | Text Encoder(s) | Params | Resolution | Steps | Year |
|---|---|---|---|---|---|---|
| SD1.5 | U-Net | CLIP ViT-L/14 | 0.86B | $512^2$ | 20 | 2022 |
| SDXL | U-Net | CLIP ViT-L/14 + OpenCLIP ViT-bigG/14 | 2.6B | $1024^2$ | 20 | 2023 |
| PixArt-$\alpha$ | DiT | T5-XXL | 0.6B | $1024^2$ | 20 | 2023 |
| Flux-dev | DiT (flow-matching) | CLIP ViT-L/14 + T5-XXL | 12B | $1024^2$ | 20 | 2024 |
| Qwen-Image | DiT (flow-matching) | Qwen2.5-VL | 20B | $1024^2$ | 50 | 2025 |
| Z-Image-Turbo | S3-DiT (distilled) | Qwen3-LM | 6B | $1024^2$ | 8 | 2025 |
| Krea 2 | MMDiT (distilled) | Qwen3-VL | 12.9B | $1024^2$ | 8 | 2026 |

COCO dataset (Lin et al., 2014) deliberately captures objects in context, and co-occurrence is key to model scene context (Zellers et al., 2018). Context is biased: models use co-occurring objects as shortcuts (Geirhos et al., 2020; Rosenfeld et al., 2018; Shetty et al., 2019) and fail on objects seen without usual companions (Singh et al., 2020). Context also matters for synthetic training data (Dvornik et al., 2018). Recent work use context to debias detectors (Cai et al., 2026; Liang et al., 2026). We show new evidence and insights into T2I objects.

**Inference time control, editing, and concept erasure**. T2I can be controlled: guidance (Dhariwal & Nichol, 2021; Ho & Salimans, 2022), negative prompts (Ban et al., 2024), interval-restricted CFG (Kynkäänniemi et al., 2024), editing prompt embeddings (Li et al., 2024; Koh et al., 2025), loss-based guidance (Bansal et al., 2024), and attention-space negative guidance (Chen et al., 2026); cross-attention manipulation (Hertz et al., 2023; Chefer et al., 2023; Rassin et al., 2023); post-hoc editing (Meng et al., 2022; Avrahami et al., 2022; Huang et al., 2025b); and detect-and-correct or best-of-$N$ search (Wu et al., 2024a; Ma et al., 2025). Concept erasure methods: training-free (Schramowski et al., 2023; Brack et al., 2023; Xiong et al., 2026), closed-form (Gandikota et al., 2024; Gong et al., 2024), and fine-tuned (Gandikota et al., 2023; Kumari et al., 2023; Srivatsan et al., 2025; Kusumba et al., 2026). We compare our control of ICOs with them.

## 3 The Unspoken

In this section, we characterize the abundance of implicit co-occurring objects (ICOs).

### 1. Setup

**Text-to-Image models**. We study seven widely-used models released 2022–2026, with distinct architectures, text encoders, training objectives, sampling steps, and resolution (Table 1).

**Generation protocols**. For each generation, we use a prompt template: "a photo of {*category*}", where {*category*} is one common object name from 80 COCO categories. We use 100 random seeds per category per model, with default hyperparameters and settings. In total, we generate 8,000 images per model, that is **56,000** images across all seven models.

**Object probes**. We use various probes to find "objects" in the generated images, spanning open set detectors and free-form captioning models: OWLv2, CLIP, RAM++, and BLIP-2. We conducted detector-agreement analysis and verified that the detections are reliable (Appendix A.2).

**Metrics**. We formally define the metrics we use in the study throughout this paper.

- *Co-occurrence rate* $\pi(c, o)$: The sample estimate of real-world joint probability that object $o$ appears in a photograph of category $c$ is: $\pi(c, o) = |\{I \in \mathcal{D} : c \in I \wedge o \in I\}|/|\{I \in \mathcal{D} : c \in I\}|$.
- *Leakage rate* $\ell(c, o)$: The rate at which a model generates object $o$ when prompted with category $c$. Estimated as the detection frequency across $n$ independently seeded generations: $\ell(c, o) = 1/n \sum_{i=1}^{n} \mathbf{1}[\text{object } o \text{ detected in image } i]$.
- *Fidelity* $\phi(c)$: The rate at which the *prompted* category $c$ itself is detected in the generated image. We can simply define it by replacing the object in $\ell$ with the prompted category: $\phi(c) = \ell(c, c)$. This is to measure the faithfulness of the generation to the prompt.
- *Excess Leakage Score* $\Delta(c, o)$: The difference between the model's generation rate and the real-world co-occurrence rate: $\Delta(c, o) = \ell(c, o) - \pi(c, o)$. $\Delta < 0$ means under-representation. Uncertainty is quantified via bootstrap 95% confidence intervals (1,000 resamples).

Table 2: Implicit Co-occurring Objects (ICOs) are abundant in all models-generated images.

| | Leak rate $\ell$ (OWLv2) | | Corroboration (%) | | | | | |
|---|---|---|---|---|---|---|---|---|
| Model | Any-obj (%) | LP-6 (%) | CLIP | RAM++ | BLIP-2 | Mean $\Delta$ | 95% CI | $\ell/\pi$ |
| Qwen-Image | 19.1 | 52.5 | 55.5 | 48.4 | 39.6 | $-0.075$ | $[-.075, -.075]$ | 0.13 |
| SD1.5 | 34.0 | 34.3 | 31.8 | 38.2 | 27.3 | $-0.068$ | $[-.068, -.067]$ | 0.22 |
| Z-Image-Turbo | 39.2 | 70.3 | 26.0 | 41.3 | 26.7 | $-0.059$ | $[-.060, -.059]$ | 0.31 |
| Krea 2 | 45.5 | 54.2 | 39.5 | 45.3 | 32.7 | $-0.056$ | $[-.057, -.056]$ | 0.35 |
| PixArt-$\alpha$ | 47.9 | 68.5 | 35.1 | 35.6 | 25.2 | $-0.053$ | $[-.054, -.053]$ | 0.38 |
| SDXL | 49.9 | 57.5 | 37.4 | 39.5 | 27.5 | $-0.051$ | $[-.052, -.050]$ | 0.41 |
| Flux-dev | 49.1 | 64.5 | 37.7 | 43.9 | 30.0 | $-0.049$ | $[-.050, -.048]$ | 0.44 |

**2. Findings**

**ICOs are abundant**. Table 2 show the average leakage rate $\ell$ for all 80 categories of seven T2I models: even the least affected model (Qwen-Image) inserts unprompted objects in 19% of the total generation, while four of the seven models are close to 50%. Look at the six most leaked object categories (LP-6), i.e., cake, sandwich, pizza, hot dog, wine glass, cup, $\ell$ climbs drastically, reaching 70% for Z-Image-Turbo. These detections are verified by independent probes (principal OWLv2, corroborators CLIP, RAM++, and BLIP-2). Each T2I model exhibits a negative mean Excess Leakage Score $\Delta$ and less than 1.0 leak-to-prior ratios $\ell/\pi$, meaning they are not inventing new co-occurrence on average. Full co-occurrence matrices are in Appendix A.14.

**Detailed prompts do not help**. An intuitive solution to reduce ICOs is to write detailed prompts for T2I, especially with newer models with better prompt-following. We design four additional prompt templates with increased complexity and check their performance. Table 3 shows the raw number of unprompted objects and the fidelity of the image regarding the prompt complexity. It turns out that adding scene detail to the prompt is not a substitute for an explicit control of ICOs. We show three representative models here, and the full table is in Appendix A.4.

**Design choices and robustness checks**. We validate the design choices and check for robustness for the findings, to ensure these are solid evidence, not artifacts by any particular model or setting. We tested multi-probe corroboration, detector agreement, threshold robustness, COCO vs. alternatives, and cross-model consistency. Details are in Appendix A.3.

## 4 THE SOURCE AND MECHANISM

Where do ICOs come from and how do T2I models generate them? In this section, we investigate the source and mechanism by looking closely at the diffusion generation process. We formulate four non-mutually exclusive hypotheses. Spearman $\rho \geq 0.3$ is considered as medium or moderate effect. We highlight key points here and the full results are in Appendix A.5.1—A.5.4.

**Hypothesis 1 (training data):** T2I models reproduce the co-occurrence patterns in training data.

1) *Co-occurrence pattern is shared in models and consistent with real-world statistics*. We first check the correlations of the 80×80 co-occurrence matrices (with four probes) of the 7 models. They have good correlation score overall at the matrix level ($\rho \approx 0.36 - 0.56, p \approx 0$), suggesting the entire pattern is shared. The correlation between models also reflects whether they share training data (such as SD1.5 and SDXL) or not (PixArt-$\alpha$). When we check their correlation with real-world data (COCO annotations), we get significant correlation ($\rho = 0.28 - 0.39$) for all seven models.

2) *Strong correlation with actual training data*. We then check the actual training data used by SD1.5 [1]. : LAION-2B (Schuhmann et al., 2022). The correlations between LAION and COCO are strong ($\rho = 0.31 - 0.39$), despite they are from different photo collections, characteristics, and independent annotations. This validates that the co-occurrence pattern presented in the general real photos is also represented in the actual training data sampled from them.

3) *Causal intervention of training data*. To go beyond the correlation analysis and confirm causality, we design a training intervention process to hijack the training distribution and see if that will affect the leakage rate of models. We use COCO's instance segmentation mask and copy-paste the target

[1]The only T2I model fully disclosed its training data.

Table 3: Prompt-complexity ladder: more detailed prompts cannot eliminate ICOs.

| | L0: bare | | L1: +object | | L2: +room | | L3: +detail | | L4: paragraph | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | #ICO | fid. | #ICO | fid. | #ICO | fid. | #ICO | fid. | #ICO | fid. |
| SD1.5 | 0.49 | 96.6 | 1.75 | 88.8 | 2.57 | 31.8 | 1.61 | 17.9 | 0.83 | 22.0 |
| Flux-dev | 1.14 | 98.1 | 4.78 | 94.1 | 2.98 | 94.1 | 2.67 | 92.6 | 2.56 | 93.0 |
| Krea 2 | 0.84 | 97.6 | 4.90 | 98.6 | 0.42 | 96.2 | 2.37 | 95.4 | 2.26 | 94.8 |

object onto a plain surface, and fine-tune a LoRA on each diffusion model. The leakage drops significantly for all models (e.g., SD1.5: $38.3\% \rightarrow 10.8\%$), with minimal fidelity lost.

**[Assessment]** Judging by the strong correlation number and effective causal intervention, training data co-occurrence is a substantial and causally confirmed contributor to unprompted object leakage.

**Hypothesis 2 (text encoder):** The embeddings in the text encoder encodes co-occurrence.

We use cosine similarity between "a photo of {prompted object}" and "a photo of {leaked object}" embeddings to check if they correlate with leaked-category frequency.

1) *CLIP vs. T5*. CLIP encoders (SD1.5, SDXL) show a small $\rho = 0.07 - 0.15$. T5 encoder alone (PixArt-$\alpha$) has minimal correlation of 0.011, not significant. Interestingly, Flux uses both T5 and CLIP, gets correlation of 0.051 that is in between. It seems a text encoder is trained with contrastive image-text training (i.e., CLIP) has stronger co-occurrence signal than a pure language one (T5).

2) *LLM encoders*. Qwen-Image and Z-Image-Turbo both use LLMs for prompt following. They show a similar weak $\rho = 0.05$. Krea 2, which uses a Qwen3-VL encoder, 0.184 correlation, is the strongest. This again suggests that joint training of vision signals in the text encoder is stronger.

3) *Falsification*. Although T5 used in PixArt-$\alpha$ contains a minimal correlation signal, the model actually leaks the same categories with similar patterns and rates as other models.

**[Assessment]** Across five structurally distinct text encoder types, we found weak yet positive correlation signals with co-occurrence. But text encoders are not the dominant source.

**Hypothesis 3 (initial latent planning):** Early in the denoising trajectory, diffusion T2I models plan multiple objects, including unprompted ones.

We probe the denoising trajectory for each model, for each identified leaked object category. At each time step, we decode the model's predicted clean-image estimation and run the OWLv2 detector to check for objects. We record the earliest time step where an ICO first becomes visually detectable. Our probing approach follows the techniques used in diffusion image editing literature (Huang et al., 2025b). Figure 2 shows the timing.

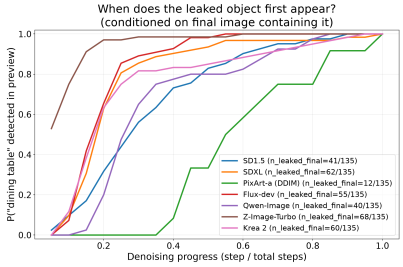


Figure 2: The first emerge timing.

1) *Six of the seven models show ICOs become detectable early*. SDXL and Flux-dev at a median of 20% of the steps, SD1.5 and Qwen-Image at 30%, and Krea 2 at 20%. PixArt-$\alpha$ is a late outlier (45%-58%). There is no clear indicator of this behavior regarding model characteristics.

2) *Z-Image-Turbo has near-instant commitment*. It generates full scene composition, including unprompted objects, at the very first denoising step. Its step-1 predicted preview is a fully composed and recognizable scene. This is not due to distillation because Krea 2 is at 20%.

**[Assessment]** ICOs are planned early (20-30% of total steps) in the latent space denoising process. Those models are trained to plant all objects in the scene from the beginning of the trajectory.

**Hypothesis 4 (sampling dynamics):** ICOs are modulated by the denoising sampling parameters.

1) *Step count and guidance scale*. More denoising steps produce more leakage across the models (up to 20 steps). Higher CFG has the same effect: a guidance scale increasing from 1.0 to 7.5, scale

Table 4: Stage 1, single control method results on reducing leakage rate. Numbers are in %.

| | SD1.5 | | SDXL | | PixArt-$\alpha$ | | Flux-dev | | Qwen-Image | | Z-Image-Turbo | | Krea 2 | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Condition | Leak | Fid. | Leak | Fid. | Leak | Fid. | Leak | Fid. | Leak | Fid. | Leak | Fid. | Leak | Fid. |
| baseline | 25.0 | 100 | 38.9 | 100 | 47.8 | 100 | 40.6 | 100 | 35.0 | 100 | 50.0 | 100 | 37.2 | 100 |
| isolation phrase | 8.9 | 98.3 | 23.3 | 100 | 10.6 | 100 | 0.0 | 86.7 | 2.3 | 99.3 | 7.8 | 99.3 | 0.0 | 83.5 |
| negative prompt | 10.2 | 98.4 | 17.2 | 98.0 | 35.9 | 100 | 2.2 | 64.4 | 33.4 | 100 | 19.0 | 97.6 | 2.8 | 95.0 |
| embedding steer | 5.6 | 95.6 | 31.7 | 100 | 42.2 | 100 | 31.1 | 96.7 | 30.0 | 93.3 | 64.2 | 100 | fails | fails |
| steps (reduced) | 7.3 | 90.1 | 3.9 | 78.6 | 15.9 | 80.6 | 25.6 | 92.8 | 28.7 | 99.1 | 47.0 | 100 | 32.2 | 100 |
| scheduled CFG | 14.4 | 98.9 | 21.1 | 97.2 | 34.4 | 100 | 1.7 | 65.6 | n/a | n/a | n/a | n/a | n/a | n/a |
| scheduler sub. | 30.4 | — | 46.9 | — | 8.9 | — | n/a | — | fails | — | fails | — | fails | — |
| detect-correct | 24.4 | 100 | 33.9 | 99.4 | 8.3 | 95.6 | 1.1 | 63.9 | n/a | n/a | n/a | n/a | n/a | n/a |
| template (best of 3) | 15.6 | 98.9 | 20.6 | 99.4 | 38.9 | 100 | 30.6 | 100 | 43.3 | 100 | 65.0 | 100 | 45.8 | 100 |

1.0 to 7.5 leads to leakage from 12-29% to 38-72%. Combined with H3, the model plans the scene with ICOs early, and additional steps and stronger CFG give it more fine-tuning to complete.

2) *Negative prompting*. Using "no dining table" is more effective on certain models such as SD1.5 (2.5x reduction) than others such as PixArt-$\alpha$ (1.3x reduction). This actually cross-validates with H2 whose text encoder showed that weak co-occurrence signal is also weak to negative prompts.

3) *Exceptions and caveats*. Qwen-Image shows category-split behavior under guidance scaling: cake, sandwich, and pizza leak more ($\rho = 0.43 - 0.48$) with higher CFG, while hot dog, wine glass, and cup leak less ($\rho = -0.30to - 0.43$). In addition, the step-distilled models Z-Image-Turbo and Krea 2 are trained to use zero guidance and low step counts.

**[Assessment]** Sampling parameters such as step count and guidance scale modulate leakage rate by amplifying or suppressing ICOs, but they do not originate the creation of them.

In summary, the hypotheses and their validation show us: T2I models derive the co-occurrence statistics from the training data, the text encoders also carry a measurable contribution. The models commit to having ICOs early in the process, and sampling parameters modulate the final results.

## 5 The Control

Can we control the unprompted objects for trained models to reduce the leakage rate of ICOs? In this section, we explore such mechanisms. In the next section, we will see there is no free lunch and having those objects is a fundamental property of text-to-image generation. We conduct a three-stage progression on the control. We focus on the single highest-frequency ICO, "dining table", as the control target for these stages, to isolate the effect of each control from variation across leak targets. Later we show generalization results across 19 leak-prone targets.

**Stage 1: single control.** We design the following individual controls inspired by Section 4.

1. *Isolation phrasing*: e.g., "with no other objects in frame, isolated on a plain white background".
2. *Negative prompting*: place the leaked category in the negative prompt and use CFG.
3. *Embedding-space steering*: project the leaked-object direction from the prompt embedding coming out from the text encoder and before it goes into the image backbone.
4. *Guidance and step reduction*: lower CFG scale and/or step count to reduce leakage.
5. *Early CFG suppression*: apply targeted negative guidance only during the early steps.
6. *Scheduler substitution*: replace the default noise scheduler.
7. *Detect and correct*: decode the preview image at first-detect timestep, inject negative guidance.
8. *Template diversity*: vary the caption template across phrasings.

Table 4 shows the results of each control on all models. The isolation phrase is the closest thing to a universal lever that reduces leakage substantially for every model and preserves fidelity above 86%. Other methods are mixed, suggesting the control can be model-dependent. Qwen-Image, Z-Image-Turbo, Krea 2 default to a guidance scale of 0, which makes sampling-based control untestable.

Table 5: Control results. **(a)** Stage 2: duo control. **(b)** Stage 3: additional controls. All values in %.

**(a) Stage 2: control duo.** Leakage / Fidelity.

| Model | Method A (alone) | Method B (alone) | Duo |
|---|---|---|---|
| SD1.5 | isolation 8.9 / 98.3 | steering 5.6 / 95.6 | **1.1 / 97.2** |
| SDXL | isolation 23.3 / 100 | negative 17.2 / 98.0 | **11.7 / 100** |
| PixArt-$\alpha$ | isolation 10.6 / 100 | negative 35.9 / 100 | **6.7 / 100** |

**(b) Stage 3: additional methods.** Leakage / Fidelity.

| Method | SD1.5 | SDXL | PixArt-$\alpha$ | Flux-dev |
|---|---|---|---|---|
| loss-guided | 2.3 / 94.0 | 4.1 / 89.2 | 1.5 / 81.5 | 4.5 / 95.1 |
| attn-suppress | 2.1 / 96.8 | 5.8 / 90.3 | 6.0 / 97.0 | — |
| post-hoc (inpaint) | 58.5 / 100 | 72.6 / 100 | — | 40.4 / 81.7 |
| post-hoc (SDEdit) | 60.4 / 100 | 85.7 / 100 | 17.5 / 95.6 | 89.4 / 100 |
| **best-of-$N$** ($N=4$) | **3.3 / 97.2** | **3.0 / 97.0** | **1.5 / 93.2** | **3.8 / 98.2** |

**Stage 2: duo control.** We further combine the two Stage 1 winners from independent mechanisms, i.e., prompt-based isolation phrase paired with either an embedding space projection or explicit negative conditioning, to achieve a control duo. With the most leakage-prone 9 categories on SD1.5, SDXL, and PixArt-$\alpha$, the combined duo beats both individual components (Table 5 (a)). The full 80-category table shows consistent results and is in Appendix A.6. We have to work with three out of the seven models so that they 1) have more than one effective individual mechanism (e.g., Flux-dev only has isolation phrase as a clean mechanism), and 2) are designed to allow for guidance scale (e.g., not Qwen-Image, Z-Image-Turbo, and Krea 2).

**Stage 3: beyond the hypotheses.** In this stage, we go beyond the hypotheses and explore additional control candidates: 1) Loss-guided sampling, using per-step gradient guidance away from the leak concept, 2) Cross-attention suppression on the unprompted tokens, 3) Post-hoc detect and edit using inpainting or 4) SDEdit, 5) best-of-N resampling. Table 5 (b) shows that best-of-N (N=4) is the only approach that transfers cleanly across all four architectures, maintaining leakage rate of 1.5-3.8%, with above 93% in fidelity. By contrast, the performance of post-hoc editing (inpainting and SDEdit) and attention suppression degrades on stronger generators.

**Comparison with concept erasure methods.** We compare against nine concept erasure methods. Table 6 shows that our hypothesis-derived control mechanism achieves competitive results. Notably, our goal is fundamentally different from concept erasure. Those methods are designed to remove a concept from the model entirely: a model has "erased" dining tables can no longer generated it even when explicitly prompted to do so. We are solving a much harder task: 1) Unprompted objects are not present in the prompt, so that token-level interventions cannot work directly. 2) Unprompted objects in one image can be prompted objects in another. Our isolation phrasing + embedding steering method is applied at inference time and suppresses the unprompted objects without changing the model's internal representation of overall object understanding. Regardless, this comparison shows that our leakage reduction is successful and has competitive performance in reducing objects in generations.

Table 6: Comparison to concept-erasure.

| Method | Leak% | Fid.% |
|---|---|---|
| Srivatsan et al. (2025) | 0.6 | 72.7 |
| Ban et al. (2024) | 1.3 | 94.8 |
| Kusumba et al. (2026) | 1.4 | 94.5 |
| Schramowski et al. (2023) | 1.5 | 95.3 |
| Gong et al. (2024) | 1.8 | 92.8 |
| Brack et al. (2023) | 2.2 | 96.1 |
| Gandikota et al. (2023) | 2.8 | 96.4 |
| Kumari et al. (2023) | 2.8 | 96.7 |
| Xiong et al. (2026) | 3.2 | 97.0 |
| Ours (iso. + steer.) | 0.19 | 94.2 |

**Generality.** To test generalization, we pick top 19 leak-prone ICOs from Section 3. We re-run the duo control against those categories as the target over all 80 prompted categories and 20 random seeds. Table A.11 in the Appendix shows that the found control approach works mostly well : e.g., person and potted plant achieve 0.0%, vase achieves 1.1% leakage, while maintaining $\approx 93\%$ fidelity. We want to point out that for the control, the only per-target input is the steering direction, therefore, it works for any leak target without retraining or architecture-specific modification. This shows that the discovered control captures a general property of a T2I model, not particular to any target object.

**Beyond binary: a continuous control.** We extend the embedding steering from one-directional suppression into a signed strength parameter $\alpha \in [-2, +2]$ to *continuously adjust* ICOs: positive values suppress, and negative values encourage them. Concretely, let $\mathbf{e} \in \mathbb{R}^d$ denote the prompt's text embedding and $\hat{\mathbf{d}}$ the unit-norm steering direction for a target category $o$. We define $\alpha \in \mathbb{R}$ and compute the steered embedding as: $\mathbf{e}' = \mathbf{e} - \alpha\,(\mathbf{e} \cdot \hat{\mathbf{d}})\,\hat{\mathbf{d}}$ when $\alpha \geq 0$ (suppress), and $\mathbf{e}' = \mathbf{e} + |\alpha|\,\hat{\mathbf{d}}$

when $\alpha < 0$ (encourage). We use two different equations because the removal operation cannot encourage an unnamed token. We test this idea by sweeping $\alpha$ over seven uniformly sampled points in $[-2, 2]$ with "dining table" as the target. Figure A.1 shows a clean, monotonic response to the parameter. More leak-prone categories have a better response with the control. This shows that the dial modulation can only affect what the model already encoded in terms of co-occurrence: for example, it won't associate a wine glass with a computer monitor. More results in Appendix A.8.

## 6 THE UNAVOIDABLE

In this section, we show that those unprompted objects are a fundamental property of large-scale photograph-trained text-to-image generative models. First, we have the following observations.

**Real photography is cluttered.** The fraction of real photographs that appear with no other categories is low (Appendix A.9), and those most-cluttered categories are also the highest leakage ones.

**Labels.** T2I models are trained with weakly-supervised text signal for the whole image. The label in text-to-image generation provides no separation between target objects and their co-occurring peers.

**Training objectives.** T2I models are trained by score matching or maximum likelihood on natural photos. For an individual image, they are trained to reconstruct every single pixel in the image, including ICOs. At the distribution level, they are a density estimator for $P_{\text{data}}(\text{image}|\text{prompt})$. A "clean" model that generates an isolated cup every time, despite cups almost never appearing isolated in the training data, would be a *worse* estimator of $P_{\text{data}}$ and discouraged during training.

**Proposition 1.** (Leakage lower bound) Let $c$ be a text prompt and $o$ be a target object. Let $\varphi_o \colon \mathcal{X} \to \{0, 1\}$ be a binary detector that returns 1 iff object $o$ is present in image $x$. Define the *data co-occurrence rate* $\pi(c, o) = \mathbb{E}_{x \sim P_{\text{data}}(\cdot|c)}[\varphi_o(x)]$ and the *model leakage rate* $\ell(c, o) = \mathbb{E}_{x \sim P_{\text{model}}(\cdot|c)}[\varphi_o(x)]$. If $D_{\text{KL}}\big(P_{\text{data}}(\cdot \mid c) \,\big\|\, P_{\text{model}}(\cdot \mid c)\big) \le \varepsilon$, then $\pi(c, o) - \ell(c, 0) \le \sqrt{\varepsilon/2}$.

*Proof* The detector $\varphi_o$ is a deterministic measurable function, so its pushforward maps each distribution to a Bernoulli: $\varphi_{o\#} P_{\text{data}}(\cdot \mid c) = \text{Bern}(\pi)$, $\varphi_{o\#} P_{\text{model}}(\cdot \mid c) = \text{Bern}(\ell)$. By the data-processing inequality, projecting both distributions through $\varphi_o$ discards all variation except the presence of $o$, so the divergence between the resulting Bernoullis can only be smaller:

$$D_{\text{KL}}\big(\text{Bern}(\pi) \,\big\|\, \text{Bern}(\ell)\big) \le D_{\text{KL}}\big(P_{\text{data}}(\cdot \mid c) \,\big\|\, P_{\text{model}}(\cdot \mid c)\big) \le \varepsilon.$$

Applying Pinsker's inequality to the left-hand side: $d_{\text{TV}}\big(\text{Bern}(\pi), \text{Bern}(\ell)\big) \le \sqrt{1/2\, D_{\text{KL}}\big(\text{Bern}(\pi) \,\big\|\, \text{Bern}(\ell)\big)} \le \sqrt{\varepsilon/2}$. For two Bernoulli distributions, $d_{\text{TV}}\big(\text{Bern}(p), \text{Bern}(q)\big) = |p - q|$, so $|\pi - \ell| \le \sqrt{\varepsilon/2}$, which gives $\pi - \ell \le \sqrt{\varepsilon/2}$. □

**Corollary 1** (Faithfulness–leakage trade-off)**.** . *For any inference-time intervention producing $P_{\text{intervene}}(\cdot \mid c)$ with leakage rate $\ell_{\text{int}} = \mathbb{E}_{x \sim P_{\text{intervene}}(\cdot|c)}[\varphi_o(x)]$,*

$$d_{\text{TV}}\big(P_{\text{data}}(\cdot \mid c),\ P_{\text{intervene}}(\cdot \mid c)\big) + \ell_{\text{int}} \ge \pi(c, o).$$

*In particular, full suppression ($\ell_{\text{int}} = 0$) makes $d_{\text{TV}} \ge \pi(c, o)$.*

*Proof.* The data-processing inequality for total variation gives $d_{\text{TV}}\big(P_{\text{data}}(\cdot \mid c),\ P_{\text{intervene}}(\cdot \mid c)\big) \ge d_{\text{TV}}\big(\text{Bern}(\pi),\ \text{Bern}(\ell_{\text{int}})\big) = |\pi - \ell_{\text{int}}|$. When $\ell_{\text{int}} \le \pi$ (the intervention reduces leakage), this becomes $d_{\text{TV}} \ge \pi - \ell_{\text{int}}$, i.e. $d_{\text{TV}} + \ell_{\text{int}} \ge \pi$. □

Proposition 1 says that any well-trained T2I model *must* reproduce the object co-occurrences present in the data. Corollary 1 further establishes a fundamental trade-off between distribution faithfulness and leakage suppression. Any post-hoc attempt to remove ICOs will either have residual co-occurrence or distort the output distribution away from the training data: it can never achieve both at the same time. The trade-off is governed by $\pi$, the co-occurrence rate in the data. This means ICOs are not an artifact or bug but a fundamental property of the real-world trained T2I models.

**Empirical $\varepsilon$ evaluation.** In practice, the bound in Proposition 1 depends on $\varepsilon$, the KL divergence between training data and model predictions, which cannot be measured directly. To assess empirical tightness, we invert the bound: given an observed leakage rate $\ell$ and the empirical co-occurrence rate $\pi$, the inequality implies $\varepsilon \ge 2(\pi - \ell)^2$. We compute this $\varepsilon$ evaluation for every *(prompt*

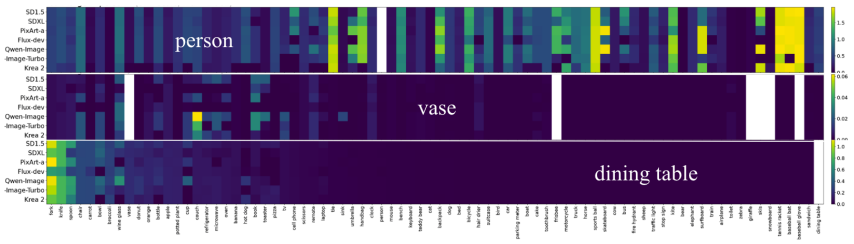


Figure 3: Estimated $\varepsilon$ for person, vase, and dining table across 7 models.

*object, leak object, model)* triple across all seven T2I models, obtaining a per-object, per-model characterization. Full results are in Appendix A.10. We found the empirical $\varepsilon$ is heavily leak-target-dependent. Figure 3 shows the estimated epsilons for person, vase, and dining table. For vase, the estimate is near zero. For dining table, $\varepsilon$ is moderate, indicating data statistics explain most but not all observed co-occurrence. For person, $\varepsilon$ is 1.9–2.0, which suggests additional leakage sources. Note that the trade-off principle (Corollary 1) holds regardless of $\varepsilon$.

**Changing $\pi$.** What if we change the training distribution? We create a separate training set of images from COCO: for each of the six leak-prone categories, we composite up to 250 photos with the object segmented out and pasted into a plain background. We then train LoRA adapters for all seven T2I models for fine-tuning to simulate the case where those models were trained on such a set. Leakage drops sharply for all models: e.g., to 0.0% for PixArt-$\alpha$, Flux-dev, Qwen, and Z-Image-Turbo with high fidelity. With $\pi_{\text{new}} \approx 0$, the T2I model can be both faithful to its (new) training distribution and free of co-occurrence leakage. More details are in Appendix A.11.

## 7 Downstream Effects

In this section, we show that downstream applications such as object detection will exhibit substantial performance gaps. **Setup**. For training data, we prepare two sets. The first is the leak-prone set with 6,000 T2I generations for 6 leak-prone categories. We use SD1.5 and Z-Image-Turbo to generate two versions of those. The second is the clean composite set with 5,400 COCO objects segment-and-pasted onto plain backgrounds. OWLv2 is used to label the bounding boxes. For test data, we use 2,408 real images from COCO, split into isolated and contextual images. For the detectors of study, we select two architecturally distinct models: Faster R-CNN and YOLO(11m). For metrics, we compute the recall drop in the test set from isolated to contextual images: the leak-prone gap is the difference in recall between the two settings for a detector trained on T2I images; the composite gap is the same difference for a detector trained on clean images. We further define gap-of-gaps: the recall drop difference between the two detectors. A positive value means ICOs amplify the context dependency, therefore affecting the performance of the detector. Recall is computed at IoU$\geq 0.25$. **Results**. Table A.16 in Appendix shows that training detectors on leak-prone categories leads to significant positive gap-of-gaps up to 18 pp for Faster RCNN and 7.7 pp for YOLO11m. This result is consistent across different diffusion models with different detectors, showing that the downstream effects from ICOs can pose a real threat to mission-critical applications.

## 8 Conclusion

We found that text-to-image models generate more objects than what they are asked for. We conducted a comprehensive study over seven architecturally distinct models, over 80 COCO categories, and 56,000 generations to study this implicit co-occurring objects (ICO) phenomenon. We characterized the rate and categories of ICOs and found a consistent pattern with real photos. We developed and verified four hypotheses about where ICOs come from, designed control mechanisms accordingly, and further showed that having ICOs is a fundamental property of T2I models trained on natural image data, such that trying to eliminate them during inference time comes with a trade-off with fidelity. We confirmed that having ICOs can lead to a drop in object detection performance in downstream tasks. Our work provides new understanding, insights, and tools to study ICOs and mitigate their impact in T2I models.

### AI USE STATEMENT

In this work, we used generative AI tools to implement methods, specifically to write the code for our generation, probing, mechanism, control, and downstream-detection experiments, and to assist in verifying and checking experimental results. The authors summarize and interpret the experiment results and findings. We used generative AI tools to initialize the theoretical understanding of the problem, and the authors derived the Proposition 1 and Corollary 1. We have not used generative AI tools to design the research methodology. Translation, dataset cleaning, and qualitative data analysis are not applicable to this work. The images produced by T2I models are the object of our study rather than AI assistance, and the generation is described in Section 3. Additionally, we used generative AI tools to maintain logs of experiment runs and polish and revise the authors' writing. We have reviewed all AI-assisted work and checked and verified the code and running results independently. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

## REFERENCES

Juveria Afreen, Mahsa Mohaghegh, and Maryam Doborjeh. Systematic literature review on bias mitigation in generative ai. *AI and Ethics*, 5(5):4789–4841, 2025.

Sumukh K. Aithal, Pratyush Maini, Zachary C. Lipton, and J. Zico Kolter. Understanding hallucinations in diffusion models through mode interpolation. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 134614–134644, 2024.

Omri Avrahami, Dani Lischinski, and Ohad Fried. Blended diffusion for text-driven editing of natural images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022.

Yuanhao Ban, Ruochen Wang, Tianyi Zhou, Minhao Cheng, Boqing Gong, and Cho-Jui Hsieh. Understanding the impact of negative prompts: When and how do they take effect? In *European Conference on Computer Vision (ECCV)*, 2024.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Roni Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Hritik Bansal, Zongyu Lin, Tianyi Xie, Zeshun Zong, Michal Yarom, Yonatan Bitton, Chenfanfu Jiang, Yizhou Sun, Kai-Wei Chang, and Aditya Grover. VideoPhy: Evaluating physical commonsense for video generation. In *International Conference on Learning Representations (ICLR)*, 2025.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwal, Casey Chu, Yunxin Jiao, and Aditya Ramesh. Improving image generation with better captions. Technical report, OpenAI, 2023. `https://cdn.openai.com/papers/dall-e-3.pdf`.

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 1493–1504, 2023.

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.

Black Forest Labs. FLUX. `https://github.com/black-forest-labs/flux`, 2024.

Manuel Brack, Felix Friedrich, Dominik Hintersdorf, Lukas Struppek, Patrick Schramowski, and Kristian Kersting. SEGA: Instructing text-to-image models using semantic guidance. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Xinhao Cai, Liulei Li, Gensheng Pei, Tao Chen, Jinshan Pan, Yazhou Yao, and Wenguan Wang. Unbiased object detection beyond frequency with visually prompted image synthesis. In *International Conference on Learning Representations (ICLR)*, 2026.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017.

Pu Cao, Feng Zhou, Qing Song, and Lu Yang. Controllable generation with text-to-image diffusion models: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025.

Hila Chefer, Yuval Alaluf, Yael Vinker, Lior Wolf, and Daniel Cohen-Or. Attend-and-excite: Attention-based semantic guidance for text-to-image diffusion models. *ACM Transactions on Graphics*, 42(4), 2023.

Dar-Yen Chen, Hmrishav Bandyopadhyay, Kai Zou, and Yi-Zhe Song. Normalized attention guidance: Universal negative guidance for diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-$\alpha$: Fast training of diffusion transformer for photorealistic text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Jaemin Cho, Abhay Zala, and Mohit Bansal. DALL-Eval: Probing the reasoning skills and social biases of text-to-image generation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3020–3031, 2023.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat GANs on image synthesis. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

Santosh K. Divvala, Derek Hoiem, James H. Hays, Alexei A. Efros, and Martial Hebert. An empirical study of context in object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1271–1278, 2009.

Nikita Dvornik, Julien Mairal, and Cordelia Schmid. Modeling visual context is key to augmenting object detection datasets. In *European Conference on Computer Vision (ECCV)*, pp. 375–391, 2018.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, et al. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Rohit Gandikota, Joanna Materzynska, Jaden Fiotto-Kaufman, and David Bau. Erasing concepts from diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Rohit Gandikota, Hadas Orgad, Yonatan Belinkov, Joanna Materzyńska, and David Bau. Unified concept editing in diffusion models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, 2024.

Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

Chao Gong, Kai Chen, Zhipeng Wei, Jingjing Chen, and Yu-Gang Jiang. Reliable and efficient concept erasure of text-to-image diffusion models. In *European Conference on Computer Vision (ECCV)*, 2024.

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *International Conference on Learning Representations (ICLR)*, 2023.

Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross-attention control. In *International Conference on Learning Representations (ICLR)*, 2023.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Han Hu, Jiayuan Gu, Zheng Zhang, Jifeng Dai, and Yichen Wei. Relation networks for object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3588–3597, 2018.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, 2025a.

Yi Huang, Jiancheng Huang, Yifan Liu, Mingfu Yan, Jiaxi Lv, Jianzhuang Liu, Wei Xiong, He Zhang, Liangliang Cao, and Shifeng Chen. Diffusion model-based image editing: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(6):4409–4437, 2025b.

Adam Tauman Kalai and Santosh S. Vempala. Calibrated language models must hallucinate. In *Proceedings of the ACM Symposium on Theory of Computing (STOC)*, pp. 160–171, 2024.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, 653(8116):1047–1051, 2026.

Eunseo Koh, Seunghoo Hong, Tae-Young Kim, Simon S. Woo, and Jae-Pil Heo. Translation of text embedding via delta vector to suppress strongly entangled content in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2025.

Nupur Kumari, Bingliang Zhang, Sheng-Yu Wang, Eli Shechtman, Richard Zhang, and Jun-Yan Zhu. Ablating concepts in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Naga Sai Abhiram Kusumba, Maitreya Patel, Kyle Min, Changhoon Kim, Chitta Baral, and Yezhou Yang. EraseFlow: Learning concept erasure policies via GFlowNet-driven alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

Tuomas Kynkäänniemi, Miika Aittala, Tero Karras, Samuli Laine, Timo Aila, and Jaakko Lehtinen. Applying guidance in a limited interval improves sample and distribution quality in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

Sangwu Lee, Erwann Millon, Le Zhuo, Matthew Newton, Andrei Filatov, Abhinay Devarinti, Dazhi Zhong, Avram Djordjevic, Gabriel Menezes, Will Beddow, Titus Ebbecke, Mihai Petrescu, Owen Fahey, Gian Saß, Felix Gil, and Victor Perez. Krea 2 technical report. `https://www.krea.ai/blog/krea-2-technical-report`, 2026.

Senmao Li, Joost van de Weijer, Taihang Hu, Fahad Shahbaz Khan, Qibin Hou, Yaxing Wang, and Jian Yang. Get what you want, not what you don't: Image content suppression for text-to-image diffusion models. In *International Conference on Learning Representations (ICLR)*, 2024.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 292–305, 2023.

Xinran Liang, Esin Tureci, Prachi Sinha, Ye Zhu, Vikram V. Ramaswamy, and Olga Russakovsky. Personalized generative models for contextual debiasing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, 2026.

Youngsun Lim, Hojun Choi, and Hyunjung Shim. Evaluating image hallucination in text-to-image generation with question-answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26290–26298, 2025.

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 *European Conference on Computer Vision (ECCV)*, pp. 740–755, 2014.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. In *International Conference on Learning Representations (ICLR)*, 2023.

Xingchao Liu, Chengyue Gong, and Qiang Liu. Flow straight and fast: Learning to generate and transfer data with rectified flow. In *International Conference on Learning Representations (ICLR)*, 2023.

Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, pp. 56338–56351, 2023.

Nanye Ma, Shangyuan Tong, Haolin Jia, Hexiang Hu, Yu-Chuan Su, Mingda Zhang, Xuan Yang, Yandong Li, Tommi Jaakkola, Xuhui Jia, and Saining Xie. Scaling inference time compute for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Chenlin Meng, Yutong He, Yang Song, Jiaming Song, Jiajun Wu, Jun-Yan Zhu, and Stefano Ermon. SDEdit: Guided image synthesis and editing with stochastic differential equations. In *International Conference on Learning Representations (ICLR)*, 2022.

Chenlin Meng, Robin Rombach, Ruiqi Gao, Diederik P. Kingma, Stefano Ermon, Jonathan Ho, and Tim Salimans. On distillation of guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Mohammad Nadeem, Shahab Saquib Sohail, Erik Cambria, Björn W. Schuller, and Amir Hussain. Gender bias in text-to-video generation models: A case study of Sora. *IEEE Intelligent Systems*, 40(3):10–15, 2025.

Aude Oliva and Antonio Torralba. The role of context in object recognition. *Trends in Cognitive Sciences*, 11(12):520–527, 2007.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, 2023.

Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. SDXL: Improving latent diffusion models for high-resolution image synthesis. In *International Conference on Learning Representations (ICLR)*, 2024.

Andrew Rabinovich, Andrea Vedaldi, Carolina Galleguillos, Eric Wiewiora, and Serge Belongie. Objects in context. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1–8, 2007.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning (ICML)*, 2021.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 2020.

Royi Rassin, Eran Hirsch, Daniel Glickman, Shauli Ravfogel, Yoav Goldberg, and Gal Chechik. Linguistic binding in diffusion models: Enhancing attribute correspondence through attention map alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 4035–4045, 2018.

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 (CVPR)*, 2022.

Amir Rosenfeld, Richard Zemel, and John K. Tsotsos. The elephant in the room. *arXiv preprint arXiv:1808.03305*, 2018.

Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. Safe latent diffusion: Mitigating inappropriate degeneration in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, et al. LAION-5B: An open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2022.

Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 6367–6384, 2024.

Rakshith Shetty, Bernt Schiele, and Mario Fritz. Not using the car to see the sidewalk – quantifying and controlling the effects of context in classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8210–8218, 2019.

Krishna Kumar Singh, Dhruv Mahajan, Kristen Grauman, Yong Jae Lee, Matt Feiszli, and Deepti Ghadiyaram. Don't judge an object by its context: Learning to overcome contextual bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11067–11075, 2020.

Koushik Srivatsan, Fahad Shamshad, Muzammal Naseer, Vishal M. Patel, and Karthik Nandakumar. STEREO: A two-stage framework for adversarially robust concept erasing from text-to-image diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2025.

Antonio Torralba. Contextual priming for object detection. *International Journal of Computer Vision*, 53(2):169–191, 2003.

Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, et al. Qwen-Image technical report. *arXiv preprint arXiv:2508.02324*, 2025.

Tsung-Han Wu, Long Lian, Joseph E. Gonzalez, Boyi Li, and Trevor Darrell. Self-correcting LLM-controlled diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2024a.

Yankun Wu, Yuta Nakashima, and Noa Garcia. Stable diffusion exposed: Gender bias from prompt to image. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, volume 7, pp. 1648–1659, 2024b.

Lexiang Xiong, Chengyu Liu, Jingwen Ye, Yan Liu, and Yuecong Xu. Semantic surgery: Zero-shot concept erasure in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

Pengfei Yang, Ngai-Man Cheung, and Xinda Ma. Text to image generation and editing: A survey. *arXiv preprint arXiv:2505.02527*, 2025.

Z-Image Team. Z-Image: An efficient image generation foundation model with single-stream diffusion transformer. *arXiv preprint arXiv:2511.22699*, 2025.

Rowan Zellers, Mark Yatskar, Sam Thomson, and Yejin Choi. Neural motifs: Scene graph parsing with global context. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5831–5840, 2018.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 2979–2989, 2017.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations (ICLR)*, 2024.

# A Appendix

## A.1 Hardware environment

All experiments ran on a single machine with 2× NVIDIA RTX PRO 6000 Blackwell Max-Q GPUs (98GB each). Software: Python 3.12.3, PyTorch (CUDA 12.8 build), diffusers 0.39.0, transformers 5.14.1 in the main environment. Root filesystem is an NVMe SSD; the working directory is on a RAID1 mirror of two SATA HDDs. All generation scripts are resume-safe (skip already-generated (category, seed) pairs on rerun), use fixed per-image seeds, and generate in batches (typically 8–16 images per batch).

## A.2 Object Probe Details

We use four different types of object probes to detect objects in the T2I-generated images.

Table A.1: Probes used to detect image content. No single extractor reliably infers what is present, so we combine four spanning detection, classification, tagging, and captioning modalities.

| Probe | Type | Vocabulary | Signal Level | Threshold |
|---|---|---|---|---|
| OWLv2 | Open-vocab detector | COCO-80 queries | Per-object, localized (boxes) | 0.15 conf. |
| CLIP | Zero-shot classifier | COCO-80 prompts | Global image-level similarity | 0.20 cosine |
| RAM++ | Open-vocab tagger | ∼4,500 tags | Multi-label, image-level | Default |
| BLIP-2 | Free-form captioner | Unconstrained | Free-text (synonym-matched) | — |

RAM++ tags and BLIP-2 captions are matched to the 80 COCO categories through a manually curated synonym table (e.g. couch/sofa, tv/television, cell phone/smartphone/mobile phone), covering all 80 categories with 0–6 additional terms each, so that wording differences do not undercount matches.

**COCO-80 categories.** person, bicycle, car, motorcycle, airplane, bus, train, truck, boat, traffic light, fire hydrant, stop sign, parking meter, bench, bird, cat, dog, horse, sheep, cow, elephant, bear, zebra, giraffe, backpack, umbrella, handbag, tie, suitcase, frisbee, skis, snowboard, sports ball, kite, baseball bat, baseball glove, skateboard, surfboard, tennis racket, bottle, wine glass, cup, fork, knife, spoon, bowl, banana, apple, sandwich, orange, broccoli, carrot, hot dog, pizza, donut, cake, chair, couch, potted plant, bed, dining table, toilet, tv, laptop, mouse, remote, keyboard, cell phone, microwave, oven, toaster, sink, refrigerator, book, clock, vase, scissors, teddy bear, hair drier, toothbrush.

## A.3 Design Choice and Robustness Validation

- *Multi-probe corroboration:* CLIP (31.8–37.7%), RAM++ (35.6–43.9%), and BLIP-2 (25.2–30.0%) all corroborate OWLv2's any-object rate (34.0–49.9%) on the same images.
- *Detector agreement:* OWLv2 and BLIP-2 in VQA mode agree at 88.3% (Cohen's $\kappa = 0.77$) on a stratified 300-image sample.
- *Threshold robustness:* a threshold-swept AUC (0.15–0.95) from the same stored per-box scores yields the identical seven-model $\ell/\pi$ ranking as the fixed 0.15 cutoff.
- $\pi$ *proxy validation:* we chose 80 COCO categories as the basis for $\pi$ because it is the gold standard for common objects in real photos. A LAION-2B training-data estimate (11.8M rows scanned, 7,429 images probed) converges with the COCO-based $\pi$: Spearman $\rho = 0.28$–$0.39$ vs. COCO, $0.31$–$0.39$ vs. LAION.
- *Cross-model consistency:* pairwise Spearman $\rho$ between models' $80\times80$ co-occurrence matrices ranges 0.36–0.51 (all $p < 10^{-190}$), confirming a shared phenomenon.

## A.4 Full Prompt Complexity Ladder

A 5-level prompt ladder, from a bare category name to a full paragraph, each level explicitly naming a superset of the objects named at the previous level: *L0* "a photo of a {category}"; *L1* adds "on

a dining table"; *L2* adds "surrounded by four wooden chairs, in a dining room"; *L3* adds "a potted plant by the window and a bowl of fruit on the table"; *L4*, a full paragraph, adds a wine glass, a vase of flowers, a clock, and a book. Table A.2 shows the full results.

Table A.2: Prompt-complexity ladder: more detailed prompts cannot eliminate ICOs.

| | L0: bare | | L1: +object | | L2: +room | | L3: +detail | | L4: paragraph | |
|---|---|---|---|---|---|---|---|---|---|---|
| Model | #unpr. | fid. | #unpr. | fid. | #unpr. | fid. | #unpr. | fid. | #unpr. | fid. |
| SD1.5 | 0.49 | 96.6 | 1.75 | 88.8 | 2.57 | 31.8 | 1.61 | 17.9 | 0.83 | 22.0 |
| SDXL | 1.03 | 97.2 | 5.80 | 94.4 | 4.25 | 73.5 | 3.74 | 47.4 | 3.98 | 61.7 |
| PixArt-$\alpha$ | 1.23 | 96.6 | 4.93 | 93.8 | 4.69 | 91.3 | 5.38 | 76.9 | 4.68 | 70.6 |
| Flux-dev | 1.14 | 98.1 | 4.78 | 94.1 | 2.98 | 94.1 | 2.67 | 92.6 | 2.56 | 93.0 |
| Z-Image-Turbo | 0.79 | 97.2 | 4.03 | 98.2 | 1.22 | 97.8 | 2.15 | 91.5 | 1.98 | 89.5 |
| Krea 2 | 0.84 | 97.6 | 4.90 | 98.6 | 0.42 | 96.2 | 2.37 | 95.4 | 2.26 | 94.8 |

## A.5 Section 4: Full Results

### A.5.1 H1: Training Data

Table A.3: Matrix-level Spearman correlation between each model's generated co-occurrence matrix (flattened, off-diagonal, OWLv2 detector) and two independent real-image co-occurrence matrices. COCO: MS-COCO 2017 train+val instance annotations (122,218 images, human-labeled ground truth). LAION: caption-keyword retrieval from `relaion2B-en-research-safe` (7,429 downloaded images, OWLv2-probed). All $p < 10^{-100}$. PixArt-$\alpha$ and Flux-dev are not LAION-trained but included as differential controls.

| Model | $\rho$ vs. COCO | $\rho$ vs. LAION |
|---|---|---|
| SD1.5 | 0.337 | 0.363 |
| SDXL | 0.394 | 0.392 |
| PixArt-$\alpha$ | 0.277 | 0.305 |
| Flux-dev | 0.388 | 0.383 |

Table A.4: Per-target H1 generalization: Spearman $\rho$ between each prompted category's real COCO co-occurrence rate with the leak target and its generated leak rate of that target (79 prompted categories per target, all 7 models). All values significant at $p < 0.02$ except where noted.

| Model | Dining table | Person | Vase |
|---|---|---|---|
| SD1.5 | 0.59–0.67 | 0.27–0.61 | 0.60–0.64 |
| SDXL | 0.59–0.67 | 0.27–0.61 | 0.60–0.64 |
| PixArt-$\alpha$ | 0.59–0.67 | 0.27 | 0.60–0.64 |
| Flux-dev | 0.59–0.67 | 0.27–0.61 | 0.60–0.64 |
| Qwen-Image | 0.477 | 0.381 | 0.180$^\dagger$ |
| Z-Image-Turbo | 0.615 | 0.636 | 0.450 |
| Krea 2 | 0.635 | 0.385 | 0.489 |

$^\dagger$ $p = 0.11$, not significant.

### A.5.2 H2: Text Encoder

For each model's own text encoder, we embedded all 80 "a photo of {category}" prompts and measured whether pairwise cosine similarity predicts detector-measured leakage frequency via Spearman rank correlation (Table A.6).

Table A.5: Top real co-occurring category for each leak-prone-6 category, from two independent real-image sources. COCO: MS-COCO 2017 train+val instance annotations ($n$ = 1,273–9,579 real images per category, human-labeled ground truth). LAION: caption-keyword retrieval from `relaion2B-en-research-safe` ($n$ = 72–115, smaller sample, noisier rates).

| | COCO | | LAION | |
|---|---|---|---|---|
| Category | Top co-occurring | Rate | Top co-occurring | Rate |
| cake | dining table | 58.4% | dining table | 20.2% |
| sandwich | dining table | 56.5% | dining table | 24.3% |
| pizza | dining table | 53.8% | dining table | 22.9% |
| hot dog | person | 56.9% | sandwich | 25.3% |
| wine glass | person | 65.3% | bottle | 21.5% |
| cup | person | 52.9% | person | 16.7% |

Table A.6: Text-encoder embedding geometry (H2): Spearman rank correlation between pairwise cosine similarity of "a photo of {category}" embeddings and detector-measured leakage frequency, per encoder across all seven models.

| Model | Encoder | $\rho$ | $p$ | Mean sim. (leaked) | Mean sim. (non-leaked) |
|---|---|---|---|---|---|
| SD1.5 | CLIP (ViT-L) | 0.102 | $4.3\times10^{-16}$ | 0.509 | 0.486 |
| SDXL | CLIP (ViT-L) | 0.071 | $1.9\times10^{-8}$ | 0.499 | 0.487 |
| SDXL | CLIP-proj (bigG) | 0.150 | $6.6\times10^{-33}$ | 0.628 | 0.609 |
| PixArt-$\alpha$ | T5-XXL | **0.011** | **0.376 (n.s.)** | 0.658 | 0.656 |
| Flux-dev | CLIP | 0.092 | $1.9\times10^{-13}$ | 0.507 | 0.487 |
| Flux-dev | T5 | 0.051 | $5.0\times10^{-5}$ | 0.670 | 0.654 |
| Qwen-Image | Qwen2.5-VL | 0.055 | $<10^{-4}$ | 0.975 | 0.972 |
| Z-Image-Turbo | Qwen3 LM | 0.052 | $<10^{-4}$ | 0.9995 | 0.9995 |
| Krea 2 | Qwen3-VL (12-layer) | **0.184** | $\mathbf{<10^{-48}}$ | — | — |

### A.5.3 H3: Initial Latent Planning

We probed each model's denoising trajectory by decoding the predicted $x_0$ at every step and running the OWLv2 detector on each preview (9-category subset, 15 seeds, 20 steps; Table A.7).

Table A.7: Denoising-trajectory probing (H3): median step fraction at which the leaked object first becomes detectable in the predicted-$x_0$ preview. Protocol: 9-category subset, 15 seeds, 20 fixed denoising steps per model.

| Model | Backbone | Conditioning | Distilled? | Median first-detect fraction |
|---|---|---|---|---|
| SD1.5 | U-Net | CLIP | No | 0.30 |
| SDXL | U-Net | CLIP + CLIP | No | 0.20 |
| Flux-dev | DiT | CLIP + T5 | No | 0.20 |
| Qwen-Image | DiT | Qwen2.5-VL LLM | No | 0.30 |
| Krea 2 | MMDiT | Qwen3-VL (12-layer) | Yes | 0.20 |
| PixArt-$\alpha$ | DiT | **T5 only** | No | 0.45 |
| **Z-Image-Turbo** | S3-DiT | Qwen3 LLM | Yes | **0.05** |

### A.5.4 H4: Sampling Dynamics

We ablated guidance scale and step count on the leak-prone-6 subset (Table A.8, A.9). For SD1.5, SDXL, and PixArt-$\alpha$, higher guidance monotonically increases leakage — CFG amplifies the full conditional mode, including typical scene context, not just the named object. Flux-dev's distilled guidance mechanism produces a noisier, non-monotonic pattern. Step count shows a consistent more-steps-more-leakage trend across all non-distilled models, plateauing around 20 steps. Negative prompting effectiveness tracks H2's text-encoder signal: SD1.5 (38%→15%), SDXL (58%→34%),

PixArt-$\alpha$ (72%→57%), Flux-dev (62.5%→3.3%). Step-distilled models (Z-Image-Turbo, Krea 2) are tested only in their native guidance-free setting.

Table A.8: Guidance-scale ablation (H4): dining-table leak rate (%) on the leak-prone-6 subset as a function of classifier-free guidance strength.

| Guidance | SD1.5 | SDXL | PixArt-$\alpha$ | Flux-dev | Qwen-Image |
|---|---|---|---|---|---|
| 1.0 | 13 | 12 | 29 | 30.0 | 42–51[†] |
| 3.0 | 32 | 45 | 57 | 74.2 | — |
| 3.5 | — | — | — | 62.5 | — |
| 5.0 | 36 | 58 | 63 | 49.2 | — |
| 7.5 (baseline) | 38 | 58 | 72 | 60.8 | — |
| 12.0 | 38 | 66 | 68 | 75.8 | — |

[†]Qwen-Image shows no clean pooled trend across `true_cfg_scale` 1.0–10.0; per-category Spearman analysis reveals cake/pizza *rise* ($\rho = 0.14$–$0.48$) while hot dog/wine glass *fall* ($\rho = -0.30$ to $-0.43$, $p < 10^{-8}$), confirmed on continuous detector confidence.

Table A.9: Step-count ablation (H4): dining-table leak rate (%) on the leak-prone-6 subset as a function of total denoising steps. Z-Image-Turbo and Krea 2 run at native guidance = 0 throughout. Krea 2 shows a genuine non-monotonic increase at 16 steps.

| Steps | SD1.5 | SDXL | PixArt-$\alpha$ | Flux-dev | Z-Image-Turbo | Krea 2 |
|---|---|---|---|---|---|---|
| 4/5 | 11 | 6 | 23 | 39.2 | 70.5 | — |
| 8 (baseline[*]) | — | — | — | — | 70.3 | 55.0 |
| 10 | 27 | 43 | 68 | 51.7 | — | — |
| 16 | — | — | — | — | — | 73.0 |
| 20 (baseline) | 38 | 58 | 72 | 62.5 | — | — |
| 28 | — | — | — | — | 66.3 | — |
| 50 | 34 | 64 | 68 | 76.7 | 62.7 | — |

[*]8 steps is the native distillation target for Z-Image-Turbo and Krea 2.

### A.6 Section 5: Stage 2 Full 80 Category Result

We validated each model's best combined intervention across all 80 COCO categories (20 seeds each; Table A.10). Full-80 leak rates are lower than the 9-category subset — the subset deliberately concentrated on leak-prone categories, making it the harder test. SD1.5 reaches 0.19% dining-table leakage at 94.2% fidelity; Flux-dev achieves complete elimination. The fidelity gap between subset and full-scale (e.g. SD1.5: 97.2%→94.2%) reflects collateral cost on categories the intervention was not tuned for, a real limitation of leak-target-specific inference-time methods.

Table A.10: Stage 2 combined-winners control at full 80-category scale (excl. dining table as prompted category). Each model's best two Stage-1 interventions, acting through independent mechanisms, are applied jointly.

| | | 9-category subset | | Full 80 categories | |
|---|---|---|---|---|---|
| Model | Method | Leak | Fidelity | Leak | Fidelity |
| SD1.5 | isolation + steering | 1.1% | 97.2% | 0.19% | 94.2% |
| SDXL | isolation + negative | 11.7% | 100.0% | 1.39% | 97.3% |
| PixArt-$\alpha$ | isolation + negative | 6.7% | 100.0% | 0.76% | 96.2% |
| Flux-dev | isolation alone | 0.0% | 86.7% | 0.00% | 90.3% |

### A.7 Section 5.5: Generality

Table A.11 shows the full results of applying the control duo method to all 19 categories. Overall, the method reduces the percentage of unprompted objects beyond the dining table category.

Table A.11: Generality: the control duo (isolation + embedding steering) applied independently to all 19 categories with $\geq 1\%$ baseline unprompted leak on SD1.5 (full-80, 20 seeds).

| Target | Base% | Ctrl% | Fid% | Target | Base% | Ctrl% | Fid% |
|---|---|---|---|---|---|---|---|
| person | 4.3 | 0.0 | 93.9 | spoon | 1.4 | 0.4 | 94.2 |
| potted plant | 3.3 | 0.0 | 93.2 | bench | 1.3 | 0.4 | 94.2 |
| vase | 3.1 | 1.1 | 94.4 | parking meter | 1.3 | 0.6 | 94.1 |
| chair | 3.1 | 0.0 | 93.6 | bowl | 1.2 | 0.4 | 94.4 |
| dining table | 2.8 | 0.2 | 93.1 | mouse | 1.2 | 1.2 | 95.4 |
| dog | 2.4 | 0.0 | 92.8 | remote | 1.1 | 0.5 | 94.9 |
| cat | 2.1 | 0.1 | 93.0 | keyboard | 1.1 | 1.0 | 94.9 |
| car | 1.8 | 0.1 | 94.2 | oven | 1.0 | 0.3 | 91.6 |
| knife | 1.5 | 0.4 | 93.3 | | | | |
| cup | 1.4 | 0.6 | 95.1 | | | | |
| **Mean** | | **0.4** | **93.8** | | | | |

### A.8 Section 5: A Continuous Control

We extended the 7-point embedding-steering sweep ($-2.0$ to $+2.0$) from the six leak-prone categories to all 79 non-self COCO-80 categories (20 seeds each, 11,060 images, SD1.5). The pooled dose-response remains monotonic but much smaller in magnitude: most categories have no natural dining-table association, so the dial has little to modulate (Table A.12). Per-category breakdown confirms that the largest dose-response ranges belong almost exactly to leak-prone-6, with a long tail at near-zero. The dial modulates associations the training distribution already encodes; it does not manufacture absent ones. The figure is at A.1.

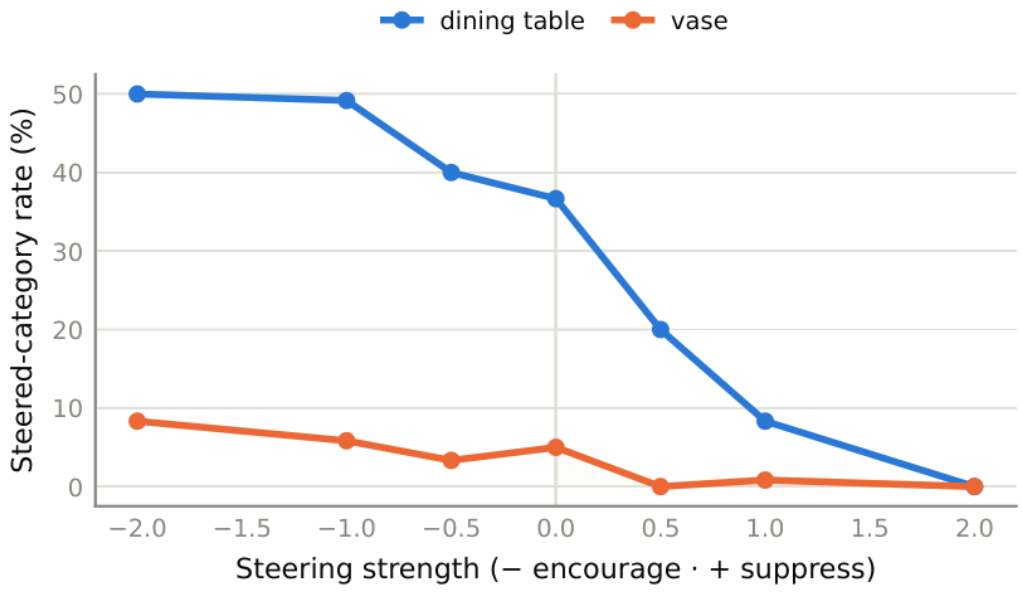


Figure A.1: Continuous dial.

Table A.12: Embedding-steering dose-response, SD1.5, dining-table target. **Top:** leak-prone-6 prompted categories ($n = 120$/point). **Bottom:** all 79 non-self COCO-80 categories ($n = 1{,}580$/point, 79 categories $\times$ 20 seeds).

| Strength | $-2.0$ | $-1.0$ | $-0.5$ | $0.0$ | $+0.5$ | $+1.0$ | $+2.0$ |
|---|---|---|---|---|---|---|---|
| *Leak-prone-6 prompted categories* | | | | | | | |
| Dining table rate (%) | 50.0 | 49.2 | 40.0 | 36.7 | 20.0 | 8.3 | 0.0 |
| Fidelity (%) | 100 | 99.2 | 100 | 98.3 | 98.3 | 93.3 | 39.2 |
| *All 79 COCO-80 categories (pooled)* | | | | | | | |
| Dining table rate (%) | 5.7 | 4.4 | 3.7 | 3.2 | 1.6 | 0.6 | 0.0 |
| Fidelity (%) | 94.4 | 95.9 | 96.5 | 96.2 | 96.0 | 93.2 | 51.4 |

### A.9 Section 6: Real Photography is Cluttered

Fraction of real COCO photographs in which the category appears with no other COCO-80 category present, from ground-truth instance annotations (122k images). Leak-prone categories are isolated 0.2–12.4% of the time in real photography; the low-leak control categories are isolated 13–75%.

Table A.13: Real photography is cluttered: 9 categories in COCO.

| Category | $n$ real images | % truly isolated |
|---|---|---|
| cake | 3,049 | 5.6% |
| sandwich | 2,463 | 5.1% |
| pizza | 3,319 | 12.4% |
| hot dog | 1,273 | 8.7% |
| wine glass | 2,643 | 0.5% |
| cup | 9,579 | 0.2% |
| zebra (low-leak) | 2,001 | 75.1% |
| banana (low-leak) | 2,346 | 13.0% |
| clock (low-leak) | 4,863 | 30.9% |

### A.10 Section 6.2: Empirical $\varepsilon$ Evaluation Full Result

Table A.14 shows the full empirical $\varepsilon$ numbers for 19 leak targets.

### A.11 Section 6.3: Changing $\pi$

We test whether changing the effective training distribution suppresses leakage directly by LoRA fine-tuning each model ($r$=8, attention projections only, 8 epochs) on 1,500 isolated-object composites built from COCO instance masks. Five of seven models reach complete elimination (Table A.15), with no per-model tuning. Fidelity is generally preserved, though the two step-distilled models (Z-Image-Turbo, Krea 2) pay a higher cost.

### A.12 Section 7: Downstream Effects

We show the full downstream object detection results in Table A.16.

Table A.14: Empirically back-solved $\varepsilon$ extended to all 19 leak-prone targets.

| Leak-target | $n$ | % holds outright | mean $(\pi - \ell)$ | median $(\pi - \ell)$ | max $(\pi - \ell)$ / implied $\varepsilon$ |
|---|---|---|---|---|---|
| person | 553 | 4.0% | +0.4507 | +0.4275 | 0.9905 / 1.9620 |
| vase | 518 | 19.7% | -0.0200 | +0.0062 | 0.1758 / 0.0618 |
| dining table | 553 | 5.4% | +0.1038 | +0.0444 | 0.7542 / 1.1376 |
| bench | 546 | 20.3% | +0.0092 | +0.0133 | 0.1713 / 0.0587 |
| knife | 497 | 8.5% | +0.0387 | +0.0094 | 0.5043 / 0.5087 |
| potted plant | 553 | 13.7% | +0.0120 | +0.0231 | 0.3820 / 0.2919 |
| dog | 553 | 1.8% | +0.0215 | +0.0191 | 0.2377 / 0.1130 |
| chair | 553 | 3.8% | +0.1321 | +0.1273 | 0.4842 / 0.4689 |
| cat | 539 | 3.9% | +0.0120 | +0.0116 | 0.1488 / 0.0443 |
| car | 553 | 5.1% | +0.0639 | +0.0265 | 0.6122 / 0.7497 |
| oven | 476 | 3.6% | +0.0379 | +0.0101 | 0.5689 / 0.6473 |
| microwave | 462 | 4.1% | +0.0265 | +0.0075 | 0.4489 / 0.4030 |
| spoon | 448 | 6.2% | +0.0424 | +0.0156 | 0.3191 / 0.2037 |
| bowl | 546 | 6.4% | +0.0757 | +0.0237 | 0.5942 / 0.7062 |
| remote | 483 | 8.7% | +0.0197 | +0.0052 | 0.3270 / 0.2138 |
| cup | 553 | 5.2% | +0.1114 | +0.0517 | 0.5301 / 0.5621 |
| parking meter | 357 | 26.9% | -0.0255 | +0.0006 | 0.0386 / 0.0030 |
| keyboard | 511 | 6.8% | +0.0112 | +0.0048 | 0.6894 / 0.9506 |
| mouse | 455 | 9.0% | +0.0123 | +0.0038 | 0.6096 / 0.7433 |

Table A.15: Training-distribution confirmation across all seven models.

| Model | Baseline leak | LoRA-composite leak | Reduction | Fidelity |
|---|---|---|---|---|
| SD1.5 | 38.3% | 10.8% | 3.5× | 100.0% |
| SDXL | 58.3% | 1.67% | 35× | 93.3% |
| PixArt-$\alpha$ | 71.7% | 0.0% | complete | 98.3% |
| Flux-dev | 64.5% | 0.0% | complete | 80.8% |
| Qwen-Image | 52.5% | 0.0% | complete | 98.3% |
| Z-Image-Turbo | 55.0% | 0.0% | complete | 80.8% |
| Krea 2 | 55.0% | 0.0% | complete | 60.0% |

Table A.16: Downstream detection bias from leak-prone synthetic training data.

| Detector | LR | Source | Leak-prone gap | Composite gap | Gap-of-gaps |
|---|---|---|---|---|---|
| YOLO11m | 0.005 | SD 1.5 | 35.5 pp | 27.8 pp | +7.7 pp |
| Faster R-CNN | 0.001 | SD 1.5 | 37.7 pp | 24.6 pp | +13.1 pp |
| Faster R-CNN | 0.005 | SD 1.5 | 39.3 pp | 21.8 pp | +17.5 pp |
| Faster R-CNN | 0.010 | SD 1.5 | 24.4 pp | 5.6 pp | +18.8 pp |
| YOLO11m | 0.005 | Z-Image-Turbo | 34.5 pp | 27.8 pp | +6.6 pp |
| Faster R-CNN | 0.001 | Z-Image-Turbo | 43.0 pp | 24.6 pp | +18.4 pp |
| Faster R-CNN | 0.005 | Z-Image-Turbo | 26.6 pp | 21.8 pp | +4.8 pp |
| Faster R-CNN | 0.010 | Z-Image-Turbo | 17.1 pp | 5.6 pp | +11.5 pp |

### A.13 VISUAL RESULTS

We show visualizations of the original generations and the attempt to control ICOs for all T2I models.

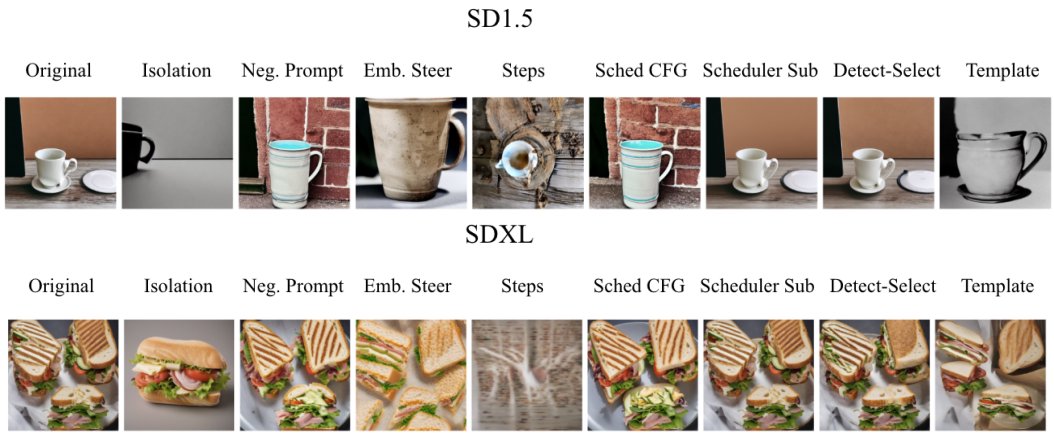


Figure A.2: Stable Diffusion 1.5 and SDXL visuals.

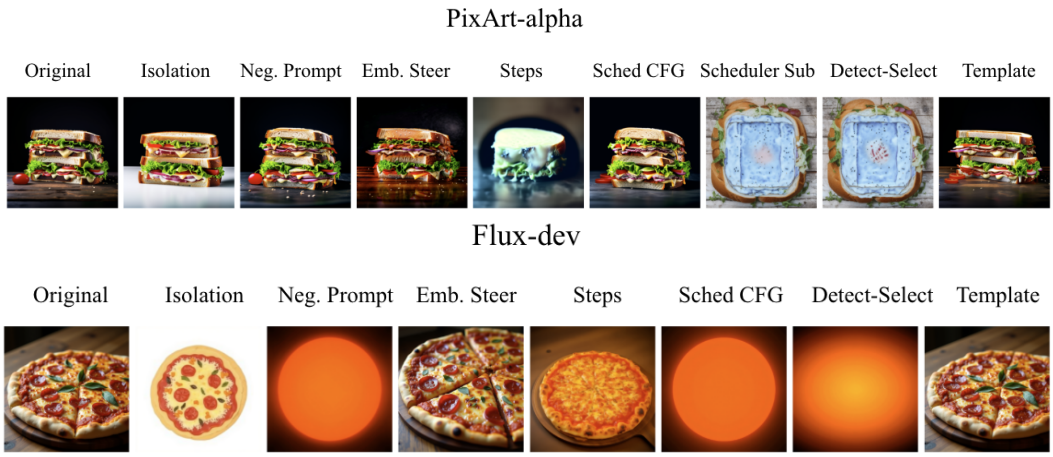


Figure A.3: PixArt-$\alpha$ and Flux visuals.

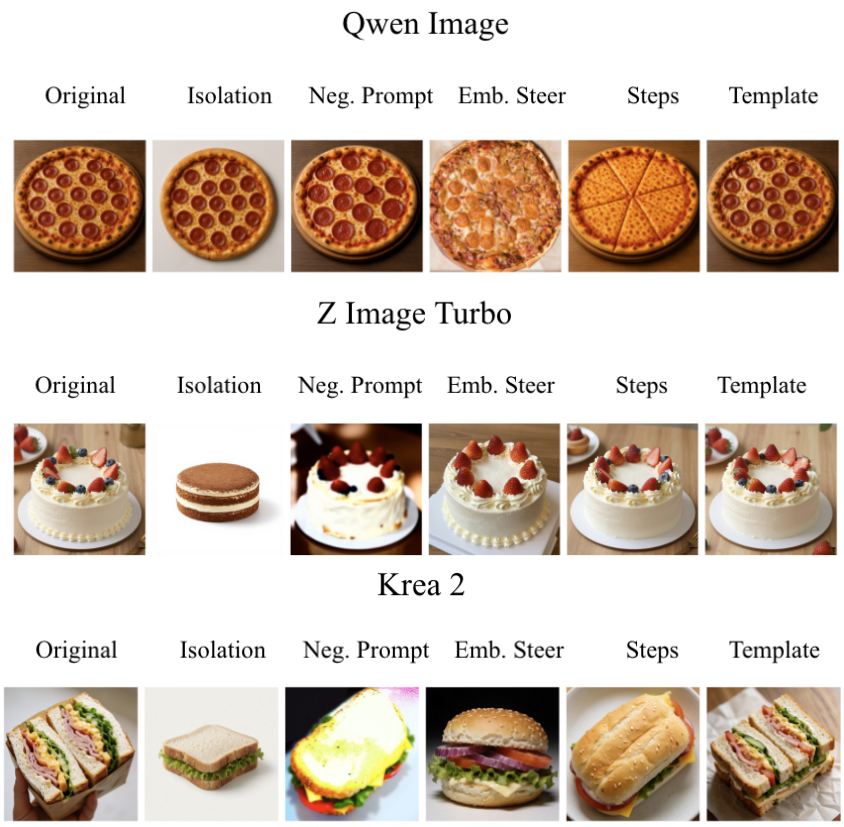


Figure A.4: Qwen Image, Z Image Turbo, and Krea 2 visuals.

### A.14 FULL CO-OCCURRENCE

In the last several pages, we show the full co-occurrence matrix for all models.

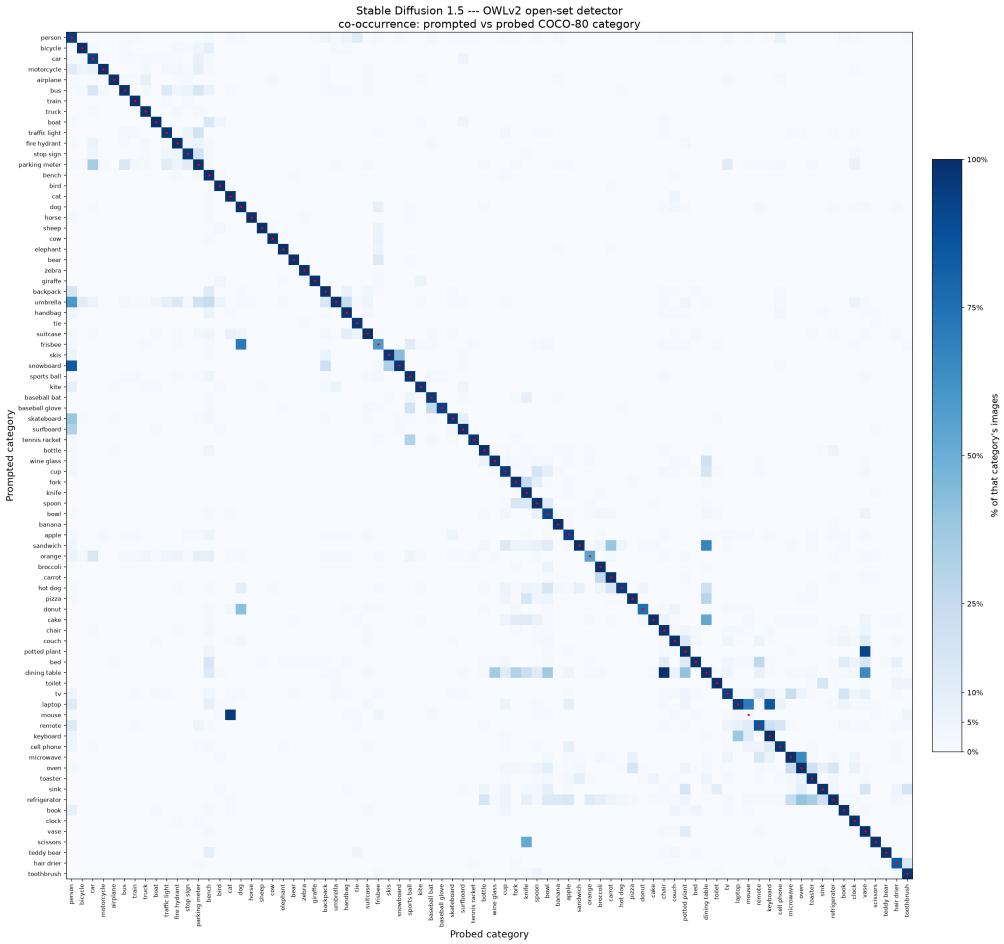


Figure A.5: Stable Diffusion 1.5 Co-Occurrence.

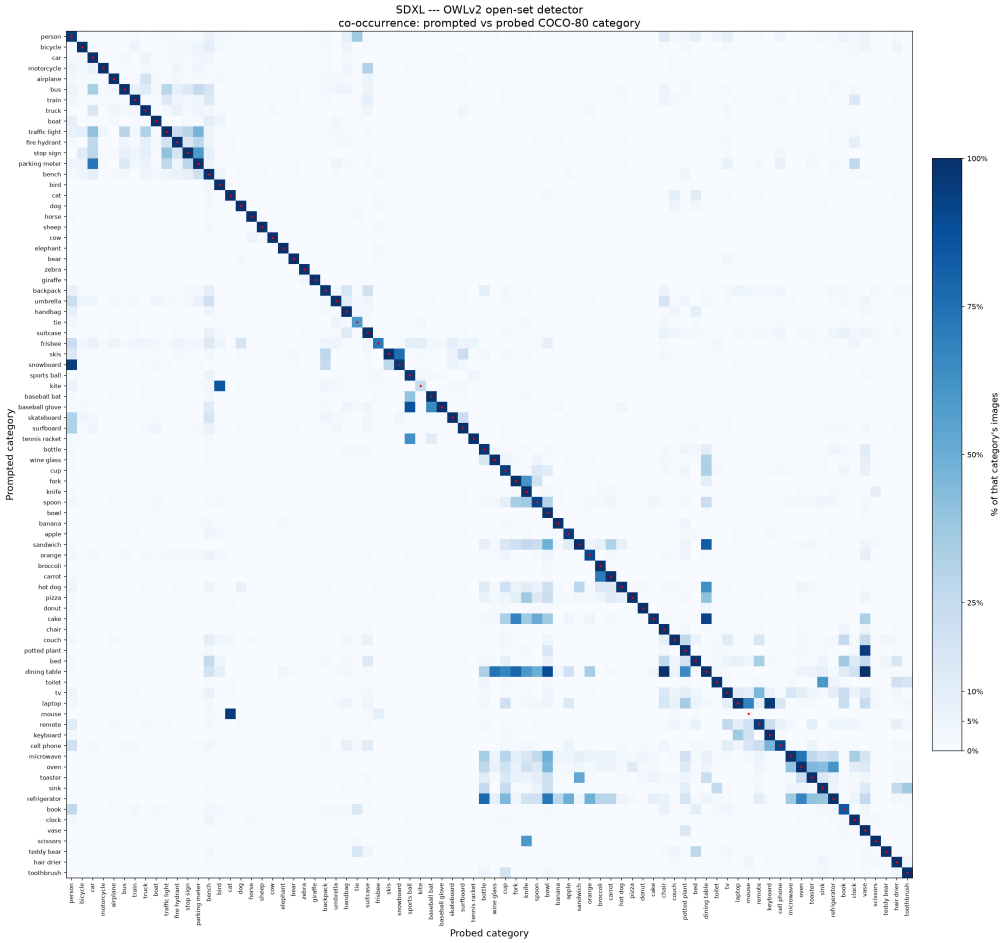


Figure A.6: SDXL Co-Occurrence.

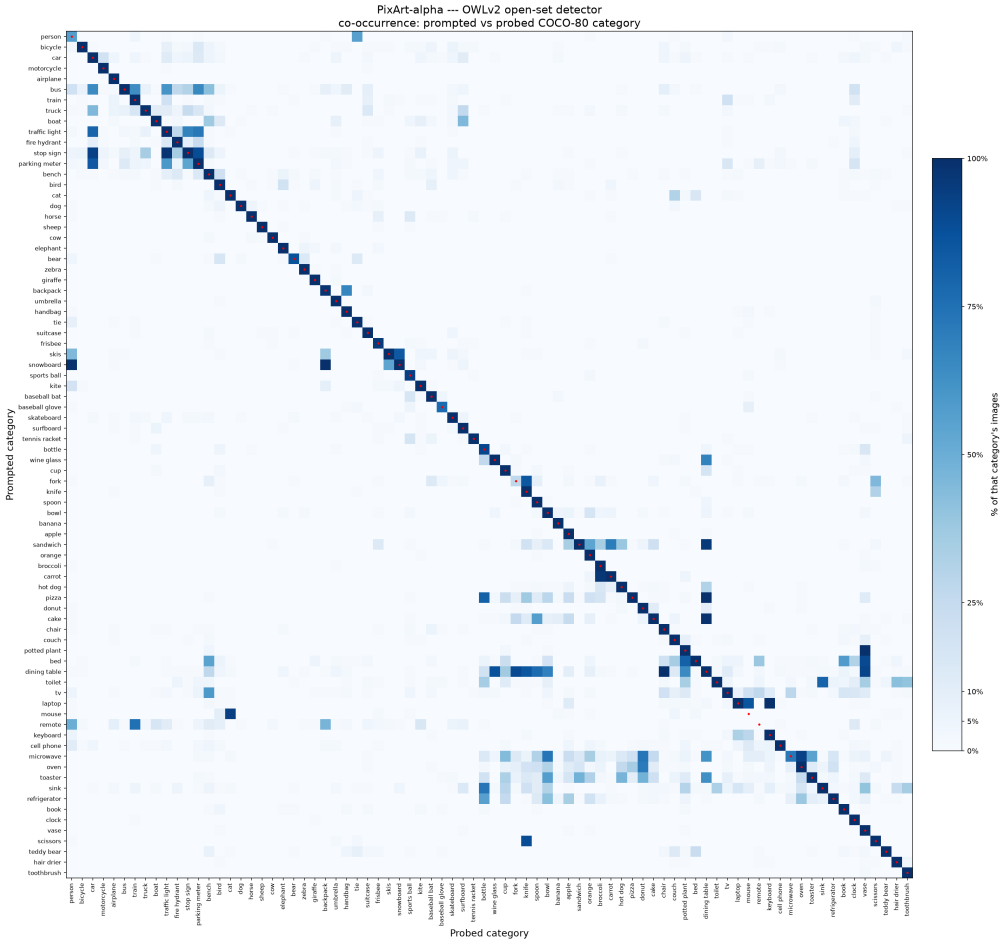


Figure A.7: PixArt-$\alpha$ Co-Occurrence.

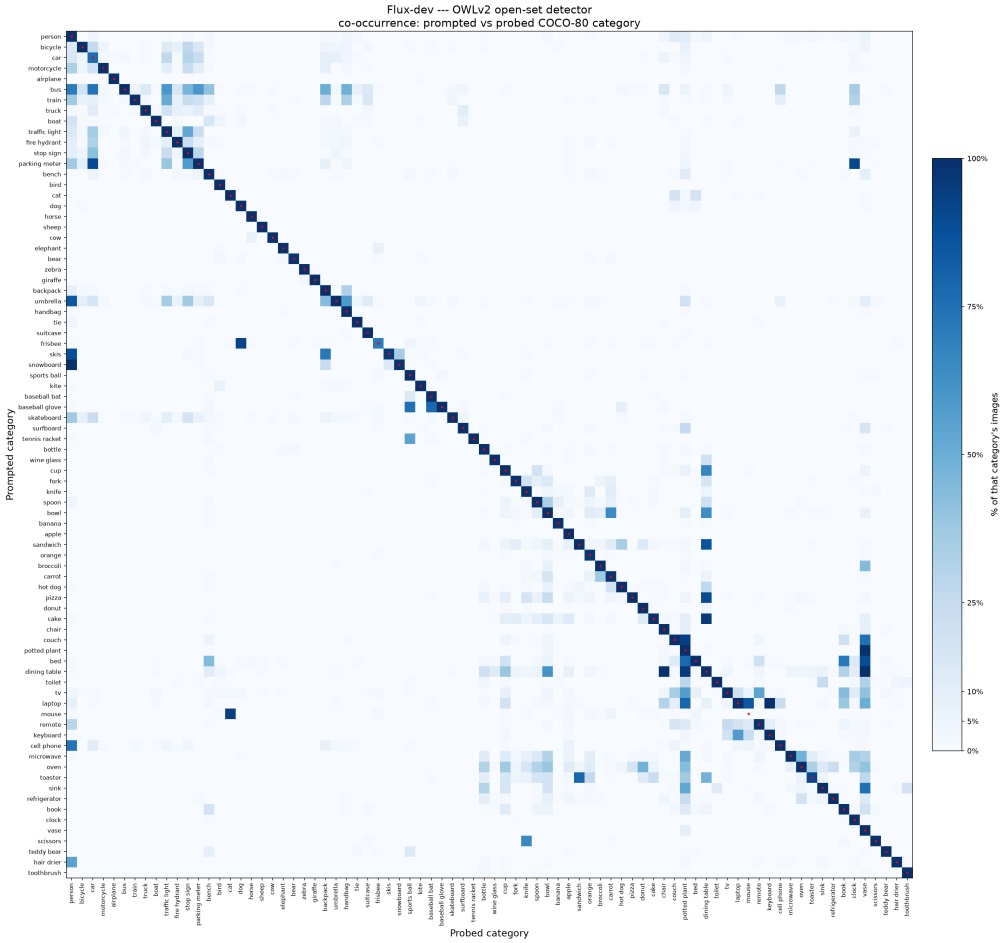


Figure A.8: Flux Dev Co-Occurrence.

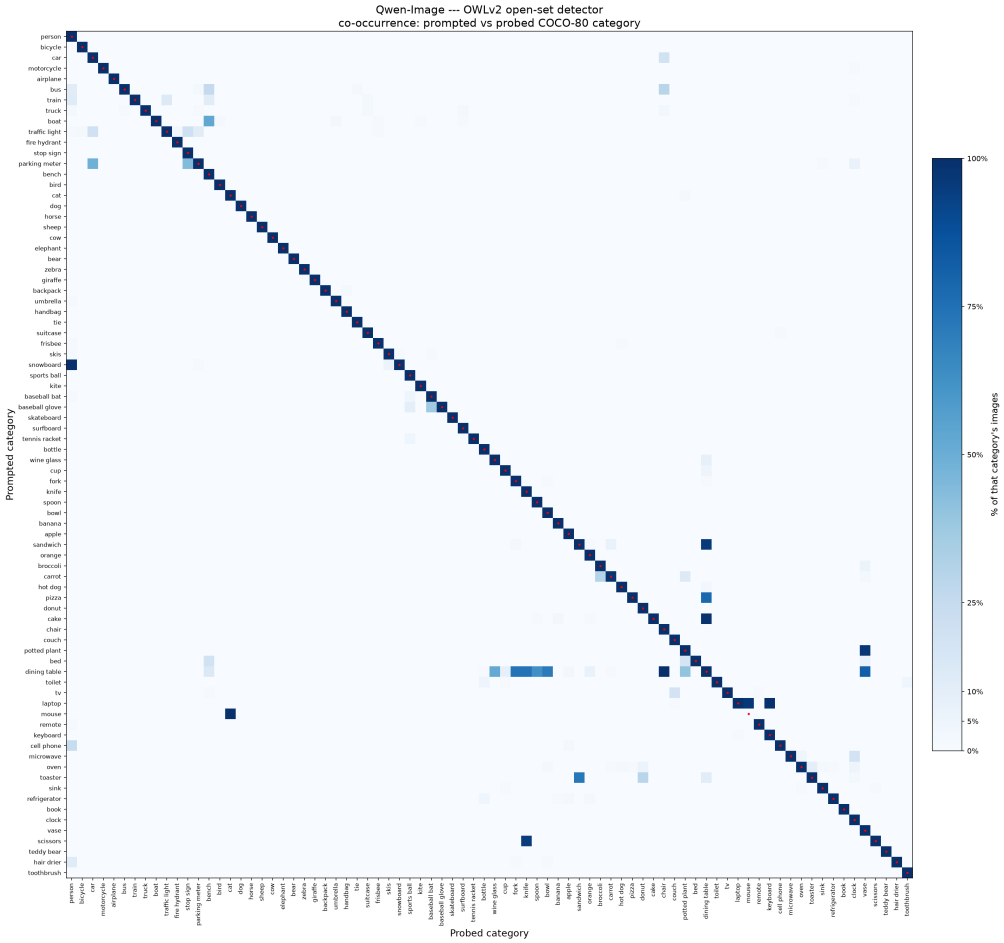


Figure A.9: Qwen Image Co-Occurrence.

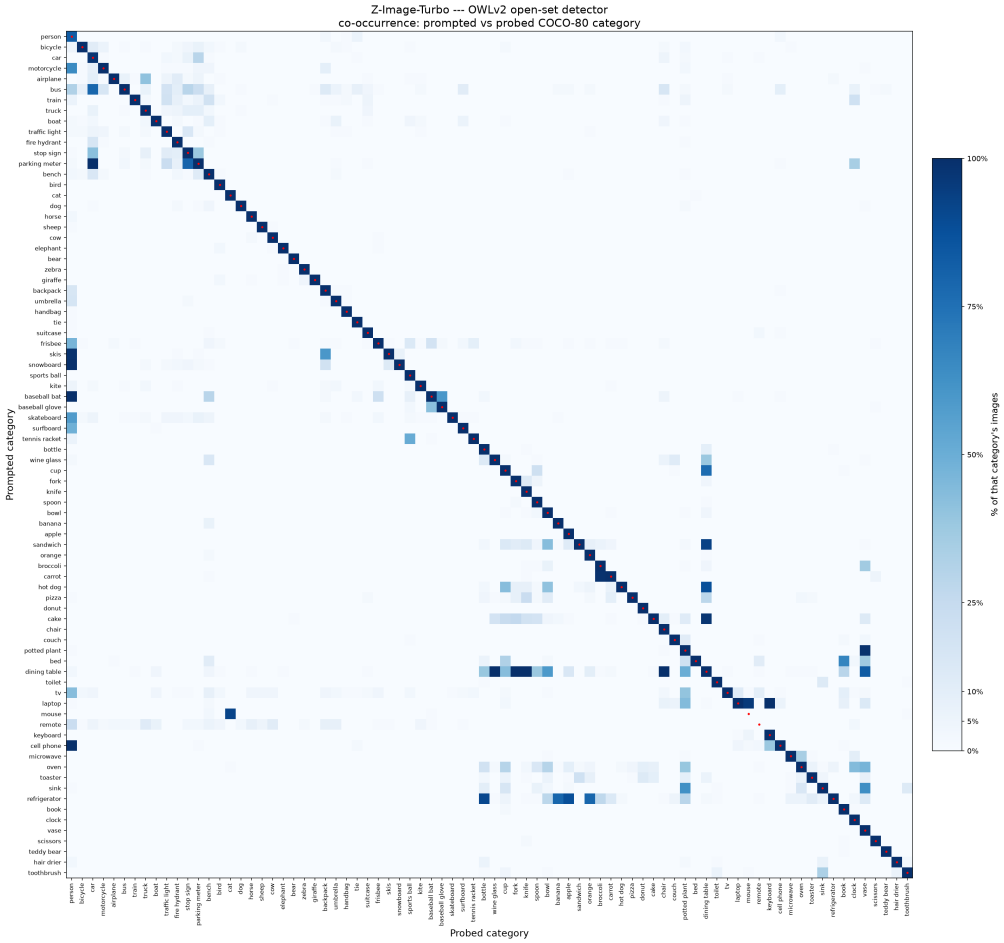


Figure A.10: Z Image Turbo Co-Occurrence.

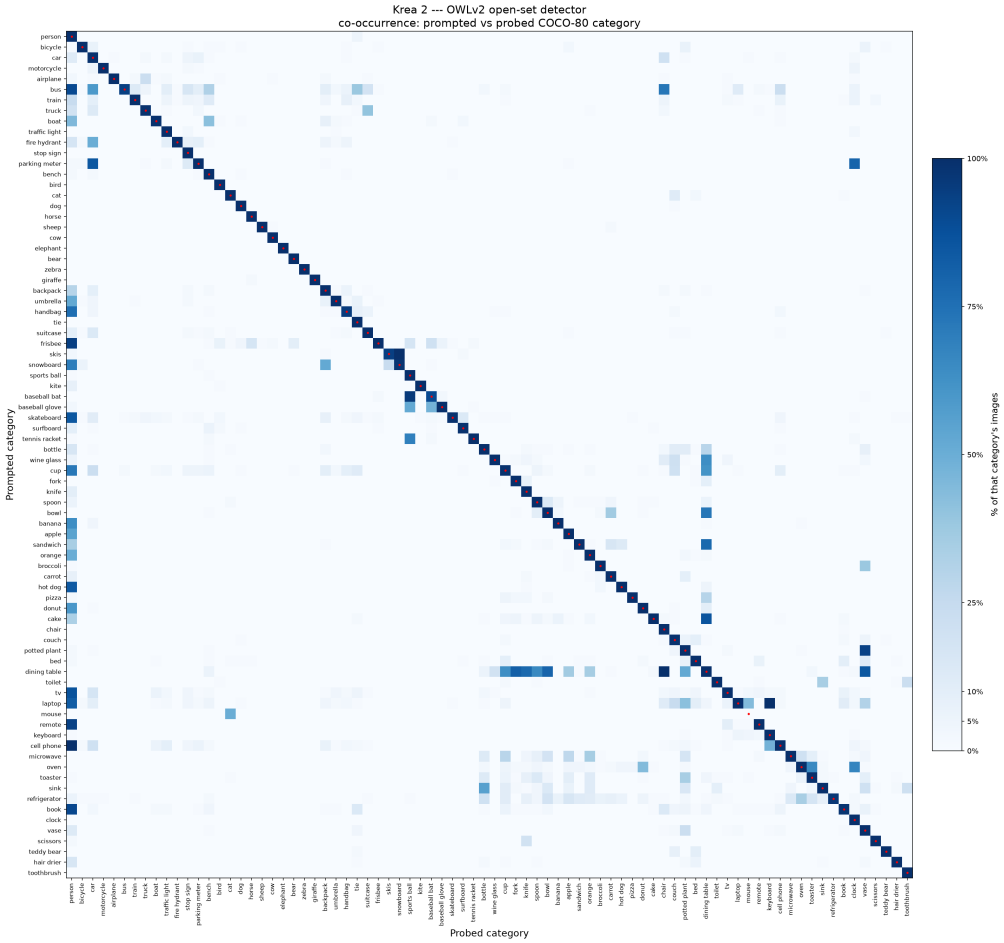


Figure A.11: Krea 2 Co-Occurrence.