COMPOSITIONAL COLLAPSE IN DIFFUSION MODELS: SCALING LAWS OF LOCAL FIDELITY

Anonymous authors
Paper under double-blind review

ABSTRACT

That diffusion models struggle with crowded scenes is known; *how* they fail — functional form, onset, and how much is measurement artifact — has never been measured. We introduce a controlled framework varying requested entity count $N \in \{1, \ldots, 64\}$ while fixing scene, lighting, and activity, measuring *local* fidelity (entity recall, facial validity, identity duplication) alongside *global* quality (CLIPScore, per-density-bin KID). Three results emerge. **(1) Collapse is exponential, not gradual:** across three open models (SDXL, SD1.5, PixArt-α), detector-corrected recall decays as $e^{-\lambda N}$ (AIC-preferred over power laws in all models, $\lambda \in [0.019, 0.034]$), giving each model a characteristic capacity scale — recall half-life of 20–36 entities — rather than scale-free degradation; the collapse reproduces with cars instead of people, with rate constants unchanged for two of the three models, indicating a largely class-independent bottleneck. **(2) Naive measurement overstates collapse by $\sim$40%:** even a strong face detector loses over a third of ground-truth faces in dense *real* scenes (recall $0.94 \rightarrow 0.59$ on WIDER FACE), a confound we calibrate against using real-image controls. **(3) Identity duplication has a measurable onset:** rare below $N \approx 8$ (0–11% of images), present in 75–84% by N=64 (ArcFace cosine > 0.4), with segmented-regression onset at $N^* \approx$ 10–22. Meanwhile global metrics are blind or anti-correlated to the collapse — CLIPScore is flat-to-increasing and KID against real crowds *improves* — revealing a decoupling between global realism and local structural correctness invisible to standard evaluation.

1 INTRODUCTION

Recent advances in diffusion-based image generation have produced systems capable of generating highly realistic images conditioned on natural language prompts (Rombach et al., 2022; Podell et al., 2023; Chen et al., 2023). These models excel in single-subject and low-density scenes. However, real-world visual scenes are inherently compositional, often involving many interacting entities with structured spatial and semantic relationships. While current models produce visually plausible global scenes, it remains unclear how their performance scales as the number of entities in a scene grows. We address the following question:

> *How does generative fidelity degrade as the number of entities in a scene increases?*

We focus on crowded human scenes as a controlled setting: people are the entity class for which the strongest instance-level measurement tools exist (person detection, face detection and alignment, identity embeddings), enabling quantitative local analysis at scale. Our hypothesis is that diffusion models exhibit systematic local failure modes under increasing entity density even when global realism remains high; Figure 1 illustrates the phenomenon.

Isn't this well known? That generative models "struggle with crowds" is indeed folk knowledge — and precisely because it is folk knowledge, it has never been characterized. Folk knowledge does not say whether fidelity degrades gracefully (a power law, with no intrinsic scale) or collapses past a characteristic capacity (an exponential, with one); whether identity failures accumulate gradually or set in over a narrow density range; or how much of the apparent failure is an artifact of running detectors in a regime where the detectors themselves fail. These distinctions carry mechanistic and practical weight: a power law and an exponential imply different capacity structures and different extrapolations, an onset density defines a safe operating range, and an uncalibrated instrument makes

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: **Compositional collapse at a glance.** The same prompt template rendered at increasing requested density N (SDXL, fixed seed, scene, lighting, and activity). Each panel magnifies one mid-scene face (red box). Global scene plausibility is maintained across the row, while local facial structure degrades with density.

all prior anecdotal claims quantitatively unreliable. Our contribution is turning a qualitative complaint into three measured, falsifiable regularities — an exponential scaling law with model-specific rate constants, a calibrated correction showing naive measurement overstates the collapse by $\sim 40\%$ at high density, and a model-dependent empirical onset of identity duplication.

A central methodological obstacle — one that, to our knowledge, prior artifact studies do not control for — is that the measurement instruments themselves degrade in crowded scenes: modern face detectors lose a large fraction of ground-truth faces as density increases even on *real* photographs. Any study that reads detector outputs on dense generated scenes at face value conflates generator failure with detector failure. We therefore calibrate every detector-based metric against real crowd imagery with ground-truth annotations, and report detector-corrected quantities.

Our contributions are as follows: 1) **An exponential scaling law for compositional fidelity.** Detector-corrected entity recall decays as $e^{-\lambda N}$ in all three models studied — exponential is AIC-selected over a power law in every case ($\lambda \in [0.019, 0.034]$) — implying a characteristic per-model capacity scale (recall half-life of 20–36 entities) rather than scale-free degradation (Section 4.2), 2) **Instrument calibration that prior artifact studies lack.** A detector-degradation control on real crowd images (WIDER FACE) shows face-detection recall falls from 0.94 to 0.59 between sparse and dense real scenes even at native detection resolution — so roughly 40% of the naively measured collapse at high density is instrument error, which we correct for (Section 3.5), 3) **An empirical onset of identity duplication.** Duplicated identities are rare below $N \approx 8$ (0–11% of images) and present in 75–84% of images at N=64, with significant segmented-regression onset estimates at $N^* \approx$ 10–22 depending on model — concentrated over a narrow density range rather than accumulating gradually (Section 4.3), 4) **A demonstration that global metrics** (CLIPScore; per-density-bin FID/KID) are flat or *improving* while local metrics collapse, quantifying the gap between global realism and local structural correctness (Section 4.4), and 5) **A controlled, released prompt framework and measurement pipeline for scene-density evaluation**, with prompt-level controls validated by pilot experiments (Section 3.1).

2 Related Work

Text-to-Image Generation. Latent diffusion (Rombach et al., 2022) and its scaled successors — SDXL (Podell et al., 2023), Imagen (Saharia et al., 2022), DALL-E 2 (Ramesh et al., 2022) — established photorealistic text-conditional generation, while transformer denoisers (Peebles & Xie, 2023; Chen et al., 2023) and rectified-flow transformers (Esser et al., 2024) now dominate the frontier. We evaluate open-weight representatives of both the UNet and DiT families; our protocol requires only prompt-conditioned samples and is therefore directly applicable to closed systems.

Evaluation of Generative Models. Standard evaluation relies on distribution-level metrics — FID (Heusel et al., 2017), Inception Score (Salimans et al., 2016), KID (Bińkowski et al., 2018) — and prompt-alignment scores such as CLIPScore (Hessel et al., 2021). Learned human-preference models (Xu et al., 2023; Kirstain et al., 2023) and QA-based faithfulness protocols (Hu et al., 2023) improve sensitivity to semantics but still operate at whole-image granularity. None of these measures instance-level structural correctness in dense scenes; we show empirically that the distribution-level metrics can *improve* while instance-level structure collapses.

Compositionality and Counting. Compositional benchmarks such as T2I-CompBench (Huang et al., 2023) and GenEval (Ghosh et al., 2023) probe attribute binding and small-count numeracy (typically $N \leq 4$), and text–image encoders are known to count poorly (Paiss et al., 2023). Interventions improve small-N fidelity through attention manipulation (Chefer et al., 2023), structured guidance (Feng et al., 2023), composable score combination (Liu et al., 2022), explicit layout grounding (Li et al., 2023), or counting-specific correction (Binyamin et al., 2024). All of this work operates at, and evaluates at, single-digit object counts. We instead measure how fidelity *scales* as

entity count grows by two orders of magnitude within a fixed scene template — the regime where we find the exponential collapse and the duplication onset.

Scaling Laws and Emergence. Scaling laws relate capability to compute, parameters, and data in language models (Kaplan et al., 2020; Hoffmann et al., 2022) and diffusion transformers (Peebles & Xie, 2023). A parallel literature debates emergent abilities — capabilities that appear abruptly with scale (Wei et al., 2022) — and whether such discontinuities are artifacts of the chosen metric (Schaeffer et al., 2023). We bring this lens to a different axis: scaling in *scene complexity* at fixed model scale. The metric-dependence critique of Schaeffer et al. (2023) directly informs our design — we report both smooth (mean similarity) and thresholded (duplicate presence) identity statistics, and we claim only a localized empirical onset, not a discontinuity.

Failure Modes, Duplication, and Memorization. Diffusion models are known to replicate training content (Carlini et al., 2023; Somepalli et al., 2023); that line of work concerns copying *from the training set*, whereas the duplication we measure is *within-image*: distinct requested entities collapsing onto a shared identity, a signature of limited concurrent identity capacity rather than memorization. Closest to our measurement target, Borji (2022) quantitatively compares the realism of *generated faces* across Stable Diffusion, Midjourney, and DALL-E 2 — but in the single-face regime, where we show all models still perform well; our contribution is measuring how face fidelity scales as faces must share a scene. Beyond that, anatomical artifacts and object duplication are widely reported anecdotally, but prior reports are essentially existence proofs: they document *that* such failures occur, at unspecified densities, measured with uncalibrated detectors. None reports the functional form of the degradation, the density range over which a failure mode sets in, or how much of the measured effect survives correcting for the instruments' own failure in crowded scenes. Our work supplies all three.

Dense Scenes in Computer Vision. Detection in crowded scenes is a long-standing challenge in its own right: tiny-face detection (Hu & Ramanan, 2017), crowd-oriented benchmarks (Yang et al., 2016; Shao et al., 2018), and density-based crowd counting (Li et al., 2018) all exist *because* standard detectors degrade under density and occlusion. This literature is precisely why generator evaluation in dense scenes cannot take detector output at face value — the motivation for our calibration protocol (Section 3.5), which appears to be absent from prior generative-model artifact studies.

3 METHOD

3.1 CONTROLLED SCENE GENERATION

We define a parametric prompt generator $\mathrm{Prompt}(N, s, l, a)$, where N is the requested number of people, s is scene type (city street, sandy beach, open-plan office), l is lighting (daylight, night/artificial), and a is activity (walking, standing and talking). We vary

$$N \in \{1, 2, 4, 8, 16, 32, 64\}, \tag{1}$$

yielding $7 \times 3 \times 2 \times 2 = 84$ prompts per model, each rendered with 5 random seeds. The template is

> "A photograph of exactly $\{N\}$ people $\{a\}$ $\{s\}$, $\{l\}$, facing the camera, no other people in the scene, realistic, highly detailed."

Terminology and scope. Compositional complexity has several distinct axes: *cardinality* (how many instances), *heterogeneity* (how varied their attributes and actions are), and *relational structure* (constraints between instances). This work isolates the cardinality axis — scene density N — while holding within-image heterogeneity fixed (all instances share one activity) and crossing scene, lighting, and activity as controlled nuisance factors. Section 4.2 verifies empirically that density, not these factors, drives the collapse; attribute binding and relational correctness at small N are studied elsewhere (Huang et al., 2023; Ghosh et al., 2023), and extending density scaling to heterogeneous and relational scenes is future work.

Two clauses are applied *uniformly across all conditions* as controls, both introduced after pilot experiments revealed specific confounds. First, "no other people in the scene": scene descriptors alone induce background pedestrians (a pilot "busy city street" prompt at $N{=}1$ produced $\sim$30 people), which would corrupt count-based metrics at low N. Second, "facing the camera": without it, only 1 of 12 pilot images at $N{=}1$ contained a detectable face (subjects rendered from behind), leaving

face-level metrics without a low-density anchor. Because both clauses are constant across N, comparisons across density remain controlled. The full design evolution, with reproduced pilot evidence, is documented in Appendix H.

3.2 MODELS

We evaluate three architecturally diverse open-weight text-to-image models: Stable Diffusion XL (Podell et al., 2023) (UNet, 1024^2, 30 steps), Stable Diffusion 1.5 (Rombach et al., 2022) (UNet, 512^2, 50 steps), and PixArt-α (Chen et al., 2023) (Diffusion Transformer, 1024^2, 20 steps), covering both convolutional-UNet and transformer denoiser families. A quality-only negative prompt ("cartoon, painting, illustration, low quality, blurry") is held constant; it deliberately does not mention faces, anatomy, or duplication, which are the quantities under measurement.

3.3 INSTANCE DETECTION AND FACE ANALYSIS

For each generated image I we run: (i) a person detector (YOLO11x (Jocher & Qiu, 2024), dense-crowd settings: 1536px inference, confidence 0.2, up to $1{,}000$ detections) giving a person count $\hat{N}_p$; (ii) a face detector with 5-point landmarks (SCRFD (Guo et al., 2021), 1024px detection resolution) giving faces $\hat{N}_f$; and (iii) ArcFace identity embeddings (Deng et al., 2019) for detected faces. Section 3.5 shows these settings matter enormously in dense scenes.

Face validity. A detected face is *structurally valid* if its landmarks satisfy basic geometric plausibility (eyes above nose above mouth, correct left–right ordering) and its sharpness is no worse than real faces *of the same size*. The sharpness test matters: Laplacian-variance blur scores depend strongly on crop size (5th-percentile scores for real WIDER FACE faces range from 1.8 at 8px to 80 at 128px), so a single global threshold conflates size with quality. We calibrate per-size-bucket thresholds at the 5th percentile of $4{,}000$ real face crops, so "validity" measures structural quality *given* size rather than penalizing smallness itself.

3.4 LOCAL FIDELITY METRICS

Entity recall. $R_p(N) = \hat{N}_p/N$ for persons and $R_f(N) = \hat{N}_f/N$ for faces. Values above 1 indicate over-generation.

Face validity rate. $V(N)$ = fraction of detected faces that are structurally valid.

Identity duplication. Identity analysis is restricted to *embeddable* faces (valid and $\geq$ 32px, below which ArcFace embeddings are unreliable). With $\hat{N}_e$ embeddable faces per image and embeddings e_i,

$$\mathrm{Dup}_\tau(N) = \frac{\#\{(i,j) : \cos(e_i, e_j) > \tau\}}{\binom{\hat{N}_e}{2}}, \qquad \hat{N}_e \geq 2, \tag{2}$$

reported for $\tau \in \{0.3, 0.4, 0.5, 0.6\}$ together with two additional statistics: the mean over faces of the maximum cosine similarity to any other face in the image (threshold-free), and the image-level *duplicate-presence* indicator $\mathbb{1}[\max_{i \neq j} \cos(e_i, e_j) > \tau]$, which, unlike the per-pair rate, is not diluted by the quadratic growth of pair count with N.

Identity probe. In the main wide-scene grid, faces at $N \geq 32$ fall below the 32px embedding floor, leaving identity analysis underpowered exactly where it matters. We therefore add a dedicated *identity probe*: a close-up group-photo prompt set ("A tightly framed group photo of exactly $\{N\}$ people standing close together, ...") over $N \in \{2, \ldots, 64\}$, two lighting conditions, and 10 seeds (120 images per model), in which faces remain large and frontal at all densities (up to $\sim$22 embeddable faces per image). All identity results are reported on this probe.

3.5 DETECTOR-DEGRADATION CONTROL

Detectors themselves fail in crowds, so raw R_f conflates generator and detector failure. We run the *identical* detection pipeline on the WIDER FACE validation set (Yang et al., 2016) — real photographs with ground-truth face boxes — binned by true face count into the same density buckets as our grid. Table 1 shows the result: matched recall (IoU > 0.5) of SCRFD on real images falls from 0.94 in sparse scenes to 0.59 in scenes of $\sim$64 faces, even with detection resolution matched to image resolution. Roughly 40% of a naively measured count "loss" at the highest density would be attributable to the instrument rather than the generator. Detection settings themselves are a further

Table 1: Detector-degradation control: the SCRFD face detector applied to *real* WIDER FACE validation images, binned by ground-truth face count. n=1,774 images, up to 300 per bin; dense-crowd detection settings (1024px detection resolution), identical to those used on generated images.

True face count bin	1	2–3	4–6	7–11	12–23	24–47	48–128
Count ratio $\hat{N}_f/N$	0.96	0.98	0.88	0.82	0.75	0.71	0.62
Matched recall (IoU>0.5)	0.88	0.94	0.86	0.80	0.72	0.68	0.59

hidden confound: at the common default configuration (640px inference), the same detectors miss $2\times$ more faces at high density, and the person detector found only 15 of $\sim$100 people in a pilot dense image versus 98 at dense-crowd settings; all our measurements use the latter, identically on generated and real images. Full per-bin control tables and the settings-sensitivity experiment are in Appendix C.

We therefore report, alongside raw face recall, the *detector-corrected* recall

$$\tilde{R}_f(N) = R_f(N) \,/\, c(N), \tag{3}$$

where $c(N)$ is the per-bin count ratio from the control. Validity and duplication are computed *conditionally on detected faces* and are therefore not subject to this correction, though we note the detected subset at high density is biased toward larger, clearer faces — making conditional metrics conservative (the true degradation can only be worse).

3.6 Object-Class Generalization Probe

To test whether the collapse is specific to human figures, we repeat the density protocol with *cars*: "A photograph of exactly $\{N\}$ cars parked [along a city street / in a large open parking lot], ..." over the same N grid (28 prompts $\times$ 5 seeds per model), counting instances with the identical dense-scene YOLO configuration. The matching real-image control uses COCO val2017 (Lin et al., 2014) images binned by ground-truth car count. Unlike faces, car detection shows *no* density degradation on real images (matched recall 0.63–0.82, non-decreasing across bins; count ratio $\approx$1), so raw generated-car recall is directly interpretable — and any collapse observed cannot be an instrument artifact. COCO val contains no images with more than $\sim$23 annotated cars, so the control covers bins up to N=16.

3.7 Global Metric Baselines

We compute CLIPScore (Hessel et al., 2021) (CLIP ViT-L/14) between each prompt and image, and per-density-bin FID and KID between generated images at each N and real WIDER FACE images of the matching density bin. KID (Bińkowski et al., 2018) is included because per-bin sample sizes make FID's small-sample bias non-negligible.

3.8 Scaling Fits and Changepoint Estimation

For each model and local metric y we fit a power law $y = cN^{-\alpha}$ and an exponential $y = ce^{-\lambda N}$ by least squares in log space over per-image data, select between them by the Akaike Information Criterion (AIC), and report bootstrap 95% confidence intervals over images. For identity metrics we fit a segmented (two-piece continuous linear in $\log_2 N$) regression; the onset density N^* is the fitted breakpoint, retained only when the segmented model beats a single line by AIC margin > 2, with bootstrap confidence intervals. This makes "scaling law" and "onset" falsifiable claims rather than descriptions of curve shape. We deliberately avoid the term *phase transition*: with seven density values on a log grid, a fitted breakpoint cannot distinguish a discontinuity from a steep sigmoid, and we claim only the weaker, well-supported statement — an empirical onset localized in N.

4 Experiments

4.1 Setup

Each of the three models generates 84 prompts $\times$ 5 seeds $=$ 420 main-grid images (1,260 total) plus 120 identity-probe images, at the native resolution and step count listed above, with fixed seeds shared across models. The complete specification (verbatim templates, factor levels, sampler and detector settings, calibration thresholds) is in Appendix A. All generation, detection, and analysis code, together with per-image manifests, will be released.

Table 2: Fitted scaling parameters (main grid; exponential $y = c\,e^{-\lambda N}$ selected by AIC over power law for all recall metrics, all models). Bootstrap 95% CIs over images. Validity: SDXL fits a slow exponential; SD 1.5 and PixArt-α are flat (decay CI includes 0).

	Person recall λ	Corrected face recall λ	Face validity
SDXL	0.011 [0.009, 0.014]	0.019 [0.014, 0.024]	λ=0.003 [0.002, 0.005]
SD 1.5	0.021 [0.018, 0.024]	0.034 [0.030, 0.039]	flat
PixArt-α	0.024 [0.022, 0.026]	0.028 [0.025, 0.032]	flat

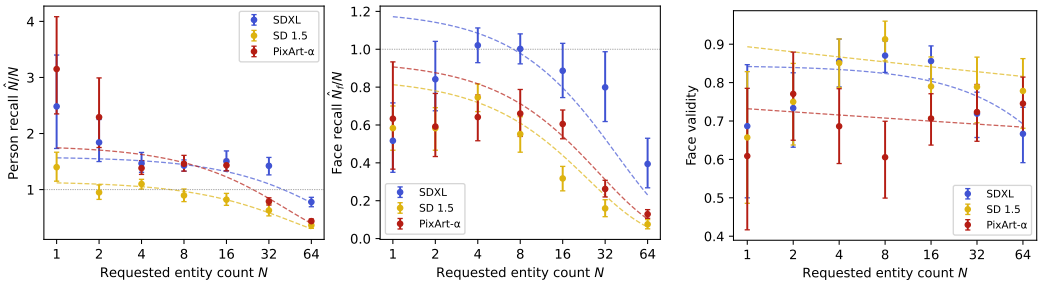


Figure 2: Left two panels: Entity recall vs. requested count N (log scale), with fitted scaling laws (dashed). Left: person recall $\hat{N}_p/N$. Right: face recall $\hat{N}_f/N$. Error bars: bootstrap 95% CIs over images. Right panel: Face validity (structural quality given size) vs. N.

4.2 Scaling of Local Fidelity

Figure 2 shows person and face recall against N, with fitted decay curves; fitted parameters are reported in Table 2, and both candidate functional forms for every metric are tabulated in Appendix B. Three findings are consistent across models. First, all models *over*-generate at low density (person recall 1.4–3.2 at N=1, background people persisting despite explicit prompt constraints) and *under*-generate at high density (recall 0.36–0.78 at N=64). Second, the decay is exponential: $y = c\,e^{-\lambda N}$ beats the power law by AIC for every recall metric in every model. Third, and critically, the decay survives detector correction — corrected face recall $\tilde{R}_f$ decays at $\lambda = 0.019$ (SDXL), 0.034 (SD 1.5), and 0.028 (PixArt-α), corresponding to recall half-lives of only 20–36 entities. Face validity (Figure 2; right panel) behaves differently: conditioned on detection it is flat for SD 1.5 and PixArt-α (CIs include zero) and declines slowly but significantly for SDXL ($0.87 \to 0.67$; λ=0.003 $[0.002, 0.005]$). The collapse is thus dominated by entities going missing or unrenderable rather than by degradation of the faces that remain detectable — noting that detection conditioning is a conservative, survivor-biased view.

Density, not the other factors, drives the collapse. Variance decomposition over the full grid shows scene choice modulates recall *levels* (e.g. $\eta^2 = 0.13$ for scene vs. 0.10 for N on SDXL person recall; for SD 1.5, N dominates at $\eta^2 = 0.25$ vs. 0.04), while lighting, activity, and seed each explain little ($\eta^2 \leq 0.09$). Crucially, the collapse *trend* is universal: in all 36 scene×lighting×activity cells across the three models, mean person recall declines from N=4 to N=64 (median ratio 0.32–0.55, maximum 0.82). Other compositional factors shift where the curve sits; density is what bends it (full decomposition in Appendix I).

4.3 Identity Collapse

Figure 3 shows identity results on the probe grid (20 images per density per model). Mean nearest-neighbor ArcFace similarity rises monotonically from ∼0.20 at N=2 to ∼0.30–0.33, saturating around N=16–32. The sharper behavior is in the image-level duplicate-presence statistic: below $N \approx 8$, SDXL produces *no* images containing a near-duplicate identity pair ($\cos > 0.4$), yet by N=64 duplicates appear in 75% of its images (84% for PixArt-α, whose onset is earlier); the pattern is robust across similarity thresholds (Appendix D). Segmented regression supports a breakpoint for SDXL at $N^* = 21.9$ $[4.5, 45.9]$ and for PixArt-α at $N^* = 10.5$ $[3.1, 17.8]$ (max-pair similarity; its main-grid breakpoints, $N^* \approx$ 6–8, are also significant). SD 1.5 shows duplicates in all images by N=32 but its estimate is not significant: at 512^2 it renders only ∼4 embeddable faces per image regardless of N, capping the probe's power. Figure 4 shows the qualitative counterpart: detected face crops shrink and lose structural coherence as density grows, with near-duplicate faces appearing at high N (a per-model duplicate gallery is in Figure 11, Appendix G). We therefore characterize identity collapse as a *model-dependent empirical onset*: duplicated identities are essentially absent at low density and near-universal at N=64, with the rise concentrated over roughly one octave of density. Whether the underlying behavior is a true discontinuity or a steep sigmoid cannot be resolved at our density resolution, and we do not claim the former.

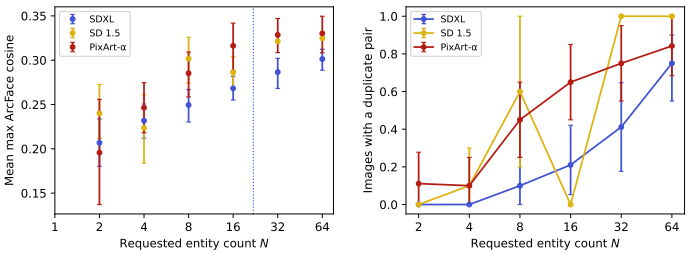


Figure 3: Identity collapse on the close-up identity-probe grid. Left: mean nearest-neighbor ArcFace cosine among embeddable faces vs. N, with dotted lines marking fitted changepoints. Right: fraction of images with at least one near-duplicate identity pair ($\cos > 0.4$) — absent at low density, near-universal by N=64.

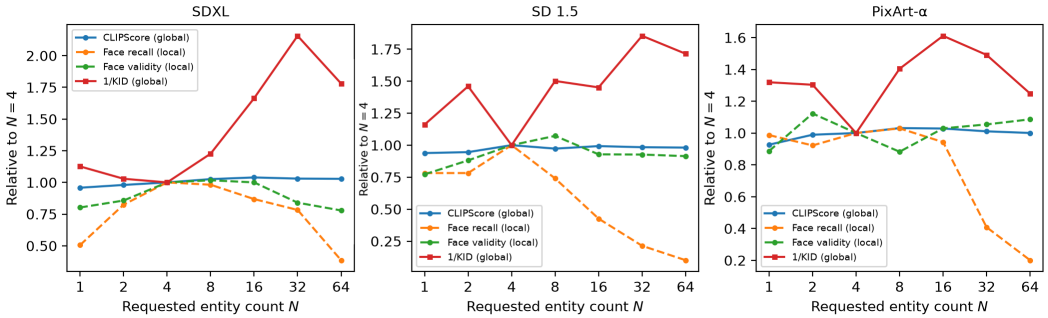


Figure 5: Global metrics stay comparatively stable while local metrics collapse. All curves normalized to their value at N=4 (per model); the N=1 anchor is avoided because it is distorted by over-generation and lone-subject face visibility.

4.4 Global vs. Local Decoupling

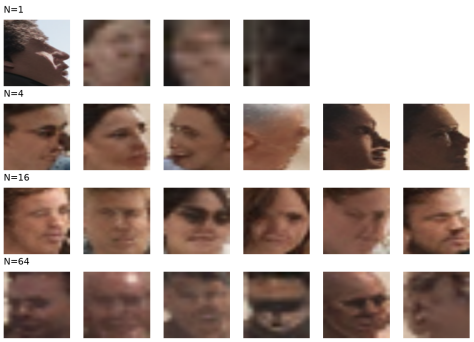


Figure 4: Detected face crops at increasing density (SDXL, seed 0). Rows: requested N. As density grows, faces shrink and structural fidelity—symmetry, eye/mouth geometry, texture coherence—visibly degrades; at high N, near-duplicate faces appear.

Figure 5 compares global metrics (CLIPScore, inverse KID) with local metrics, each normalized to its N=4 value (the N=1 anchor is distorted by over-generation and lone-subject face visibility). The decoupling is stark. Over the range where detector-corrected face recall falls 2–7× and duplicate identities emerge in most images, CLIPScore is flat to mildly *increasing* for all three models (SDXL 27.2 → 29.2; SD 1.5 25.7 → 26.9; PixArt-α 26.7 → 28.8), and per-bin KID against real crowd images of matched density *improves* or stays flat (SDXL 0.127 → 0.066–0.080; SD 1.5 0.088 → 0.055–0.060; PixArt-α flat at ∼0.13). Dense generated scenes are, distributionally, *closer* to real crowds than sparse generated scenes are to real sparse photos — at precisely the densities where instance-level structure collapses. Global metrics are not merely insensitive to compositional failure; they can be anti-correlated with it. (Absolute FID/KID values reflect the style gap between generated imagery and WIDER FACE photographs; the per-bin trend is the informative quantity.)

Effect of detector correction. Figure 6 (left) shows the detector-degradation control and its effect on the estimated collapse. Raw recall declines more steeply than detector-corrected recall because the detector increasingly misses faces in dense scenes, even on real photographs. Applying density-matched WIDER FACE correction raises recall at high N but does not alter the main result: recall still declines approximately exponentially with scene density. The collapse is therefore not a detector artifact, though naive measurements substantially overestimate its magnitude at high density.

4.5 Object-Class Generalization: Cars

The collapse is not about faces. On the car grid (Section 3.6), all three models over-generate at low density (recall 3.3–4.7 at N=1) and collapse at high density (recall 0.10–0.92 at N=64), while CLIPScore again *rises* with N for every model (Figures 7 & 6 right panel). For SDXL and PixArt-α the exponential form is again AIC-selected, and — strikingly — the fitted rates match each model's face-recall rates within confidence intervals: SDXL $\lambda_{\text{cars}} = 0.018\ [0.012, 0.023]$ vs. $\lambda_{\text{faces}} = 0.019\ [0.014, 0.024]$; PixArt-$\alpha$ $\lambda_{\text{cars}} = 0.027\ [0.022, 0.031]$ vs. $\lambda_{\text{faces}} = 0.028\ [0.025, 0.032]$. The per-

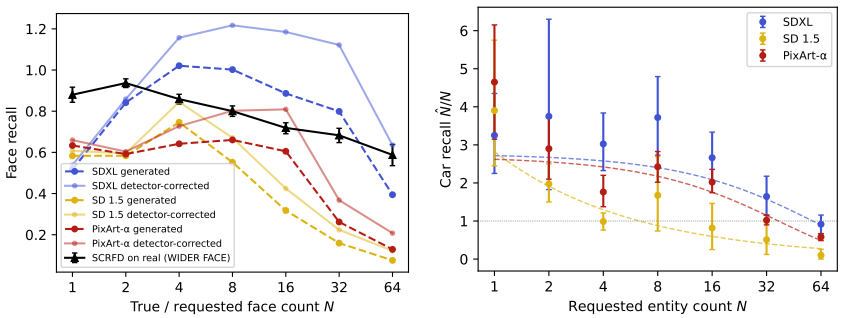


Figure 6: Left: Detector control: SCRFD face recall on real WIDER FACE images (black) vs. raw and detector-corrected face recall on generated images. Right: Car recall vs. requested count N with fitted decay curves (dashed). SDXL and PixArt-α are AIC-selected exponential with rates matching their face-recall rates; SD 1.5 is better fit by a power law. Error bars: bootstrap 95% CIs.

entity decay rate appears to be a *model-level constant across entity classes*, exactly what a shared capacity bottleneck (Section 5.3) predicts and what a face-specific explanation does not. SD 1.5 deviates: its car recall is better fit by a power law ($\alpha = 0.56$), driven by extreme over-generation at N=1 (3.9×); its collapse from N=2 onward remains monotone (1.98 → 0.10). Because the COCO control shows no detector degradation for cars, none of this can be attributed to the instrument.

5 Discussion

5.1 Compositional Collapse Regime

Our results suggest diffusion models operate in two regimes: a low-density regime of *over*-generation with high local fidelity and no identity duplication, and a high-density regime of under-generation, missing entities, and near-universal identity duplication. For identity consistency the boundary is a measurable, model-dependent onset density N^*; for recall the transition is smooth and exponential. The failure mode conditioned on detection is also model-dependent — SDXL's detected faces degrade structurally with density while SD 1.5's and PixArt-α's do not — suggesting different models spend their limited per-entity capacity differently.

5.2 Why the Functional Form Matters

The exponential-versus-power-law distinction is not curve-fitting pedantry. A power law $N^{-\alpha}$ is scale-free: degradation is gradual everywhere and no density is special. An exponential $e^{-\lambda N}$ has a characteristic scale $1/\lambda$: each additional entity multiplies expected recall by a constant factor, as if every entity independently survives generation with fixed probability — a constant per-entity hazard. That AIC selects the exponential for every recall metric in every model is evidence that current diffusion models have an intrinsic, finite *entity budget* (half-lives of 20–36 people), not merely a soft preference for simpler scenes. The two forms also extrapolate very differently: under the fitted exponentials, recall at N=200 is essentially zero for all three models, whereas a power law would predict usable partial recall. Capacity-scale estimates like $1/\lambda$ give model developers a concrete, comparable number to report and improve.

5.3 Mechanisms: Candidates and Evidence

The most informative observation is that the collapse is *architecturally invariant*: a convolutional UNet (SD 1.5), a larger UNet with dual text encoders (SDXL), and a Diffusion Transformer conditioned on T5 (PixArt-α) exhibit the same exponential form and duplication onset despite sharing neither denoiser architecture nor text encoder. This argues against explanations tied to a specific attention layout or context length, and points instead to what the three systems share:

(i) A count-blind conditioning interface. We probed SDXL's cross-attention maps for the counting-noun token (averaged over heads and layers, second half of denoising, 21 instrumented generations). The token's attention is statistically indistinguishable from uniform at *every* density (spatial entropy 6.930 vs. 6.931 for uniform; attention mass inside detected person boxes matches person pixel area within 2–13%). There is no per-instance allocation to crowd: the collective noun provides a spatially undifferentiated scene-level signal at all N, arguing against cross-attention dilution as the driver and pointing the bottleneck downstream, at the latent canvas. (Probe details, full metrics, and map visualizations: Appendix E.)

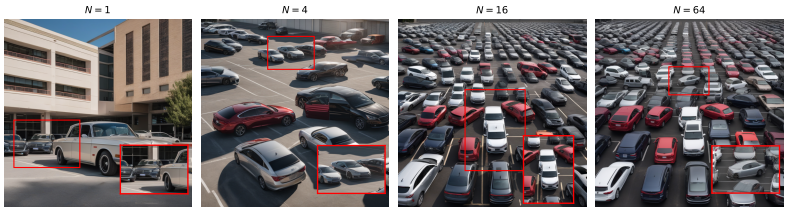


Figure 7: **The collapse generalizes beyond faces.** "Exactly N cars parked in a large open parking lot" (SDXL, fixed seed and lighting). N=1 already shows over-generation; by N=16–64 the model produces unbounded lots in which neighboring cars merge and lose geometric integrity (red insets), while each panel remains globally plausible.

(ii) A fixed latent canvas divided among entities. All three models generate in a spatially compressed latent ($8\times$ downsampling). At N=64 in a 1024^2 image, each person occupies $\sim$12 latent pixels and each face fewer than 4 — per-entity capacity shrinks as $1/N$ while the canvas stays fixed, directly predicting a constant per-entity hazard. Fitted rates track this: decay is slowest for the largest canvas (SDXL, λ=0.019 at 1024^2), faster for PixArt-α (0.028 at 1024^2 with a smaller transformer), and fastest for SD 1.5 (0.034 at 512^2, one quarter the canvas).

(iii) Count-impoverished training supervision. Web captions rarely state accurate counts beyond small numbers, and text encoders are demonstrably weak counters; no model has seen dense supervision pairing "exactly N" with veridical N at large N. This predicts both the over-generation at low N and the under-generation at high N: the model has learned "crowd" as a texture-like attribute rather than a countable set.

(iv) Identity collisions under finite concurrent capacity. The duplication onset resembles a birthday-collision phenomenon: if an image draws identities from an effective pool of size K, collision probability rises sigmoidally in N with no discontinuous mechanism required. Fitting this one-parameter model against a two-parameter logistic by AIC, SDXL favors the mechanistic model decisively ($\Delta\text{AIC} = 15.8$), estimating a pool $K \approx 2{,}200$ [1,244, 3,576]; SD 1.5 weakly agrees ($K \approx 196$). PixArt-α deviates (logistic preferred, $\Delta\text{AIC} = 3.3$): duplication is already elevated at N=2, suggesting an additional density-independent mechanism, consistent with its smaller pool ($K \approx 92$) and earlier onset. (Model derivation and fit overlays: Appendix F.)

Three independent measurements point to the latent-canvas account (ii): the attention probe shows conditioning is spatially uniform (so resources aren't allocated in cross-attention), the decay constant is invariant across entity classes ($\lambda_{\text{cars}} \approx \lambda_{\text{faces}}$ per model), and λ orders by canvas size across models. The candidates remain discriminable: a less-compressed latent should reduce λ if (ii) dominates; instance-grounded conditioning tests (i); fine-tuning on count-verified captions tests (iii).

5.4 Evaluation Implications

Standard metrics fail to capture compositional failure modes. Our detector-degradation control shows instance-aware evaluation is itself nontrivial — the instruments degrade exactly where the phenomenon of interest lives, and future compositional benchmarks should calibrate against real-image controls as we do here.

6 Limitations

Five limitations apply: the detector-correction assumes generated and real crowds stress detectors equivalently, leaving possible residual bias; identity analysis at high density excludes small faces below ArcFace's $\sim$32px resolution floor; requested count N is a soft target since models add background people even at low N; the seven-point log-density grid can characterize decay form and localize onsets but cannot distinguish a sharp discontinuity from a steep sigmoid; and scope is limited to human/car scenes across three open models at native resolution, with proprietary systems and other entity classes untested.

7 Conclusion

We studied how diffusion models behave under increasing scene density with a detector-calibrated measurement pipeline. Detector-corrected entity recall decays exponentially with density in all three models studied (half-lives of 20–36 entities); duplicated identities emerge beyond a model-dependent onset density and appear in 75–84% of images at N=64; and global perceptual metrics are not merely blind to this collapse but can improve as it occurs. These findings position compositional density as a first-class axis for evaluating — and eventually training — generative models.

AI USE STATEMENT

In this work, we used generative AI tools for editing the manuscript for readability, formatting references, and drafting/editing code. We have not used generative AI tools to generate synthetic data, formulate mathematical claims, or write proofs. We also used generative AI tools to draft parts of the text and summarize the related literature. All AI-generated code and text were reviewed and verified by the authors. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REFERENCES

Mikołaj Bińkowski, Danica J Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations (ICLR)*, 2018.

Lital Binyamin, Yoad Tewel, Hilit Segev, Eran Hirsch, Royi Rassin, and Gal Chechik. Make it count: Text-to-image generation with an accurate number of objects. *arXiv preprint arXiv:2406.10210*, 2024.

Ali Borji. Generated faces in the wild: Quantitative comparison of Stable Diffusion, Midjourney and DALL-E 2. *arXiv preprint arXiv:2210.00586*, 2022.

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *32nd USENIX Security Symposium*, pp. 5253–5270, 2023.

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 (TOG)*, 42(4):1–10, 2023.

Junsong Chen, Jincheng Yu, Chongjian Ge, Lewei Yao, Enze Xie, Yue Wu, Zhongdao Wang, James Kwok, Ping Luo, Huchuan Lu, and Zhenguo Li. PixArt-α: Fast training of diffusion transformer for photorealistic text-to-image synthesis. *arXiv preprint arXiv:2310.00426*, 2023.

Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 4690–4699, 2019.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis. In *International Conference on Machine Learning (ICML)*, 2024.

Weixi Feng, Xuehai He, Tsu-Jui Fu, Varun Jampani, Arjun Akula, Pradyumna Narayana, Sugato Basu, Xin Eric Wang, and William Yang Wang. Training-free structured diffusion guidance for compositional text-to-image synthesis. In *International Conference on Learning Representations (ICLR)*, 2023.

Dhruba Ghosh, Hannaneh Hajishirzi, and Ludwig Schmidt. GenEval: An object-focused framework for evaluating text-to-image alignment. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Jia Guo, Jiankang Deng, Alexandros Lattas, and Stefanos Zafeiriou. Sample and computation redistribution for efficient face detection. *arXiv preprint arXiv:2105.04714*, 2021.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. CLIPScore: A reference-free evaluation metric for image captioning. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 7514–7528, 2021.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs trained by a two time-scale update rule converge to a local Nash equilibrium. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 30, 2017.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Peiyun Hu and Deva Ramanan. Finding tiny faces. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 951–959, 2017.

Yushi Hu, Benlin Liu, Jungo Kasai, Yizhong Wang, Mari Ostendorf, Ranjay Krishna, and Noah A Smith. TIFA: Accurate and interpretable text-to-image faithfulness evaluation with question answering. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20406–20417, 2023.

Kaiyi Huang, Kaiyue Sun, Enze Xie, Zhenguo Li, and Xihui Liu. T2I-CompBench: A comprehensive benchmark for open-world compositional text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Glenn Jocher and Jing Qiu. Ultralytics YOLO11. `https://github.com/ultralytics/ultralytics`, 2024.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

Yuval Kirstain, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. Pick-a-pic: An open dataset of user preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Yuheng Li, Haotian Liu, Qingyang Wu, Fangzhou Mu, Jianwei Yang, Jianfeng Gao, Chunyuan Li, and Yong Jae Lee. GLIGEN: Open-set grounded text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 22511–22521, 2023.

Yuhong Li, Xiaofan Zhang, and Deming Chen. CSRNet: Dilated convolutional neural networks for understanding the highly congested scenes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1091–1100, 2018.

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.

Nan Liu, Shuang Li, Yilun Du, Antonio Torralba, and Joshua B Tenenbaum. Compositional visual generation with composable diffusion models. In *European Conference on Computer Vision (ECCV)*, pp. 423–439, 2022.

Roni Paiss, Ariel Ephrat, Omer Tov, Shiran Zada, Inbar Mosseri, Michal Irani, and Tali Dekel. Teaching CLIP to count to ten. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3170–3180, 2023.

William Peebles and Saining Xie. Scalable diffusion models with transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 4195–4205, 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. *arXiv preprint arXiv:2307.01952*, 2023.

Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with CLIP latents. *arXiv preprint arXiv:2204.06125*, 2022.

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)*, pp. 10684–10695, 2022.

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

Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, and Xi Chen. Improved techniques for training GANs. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, 2016.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. Are emergent abilities of large language models a mirage? In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shuai Shao, Zijian Zhao, Boxun Li, Tete Xiao, Gang Yu, Xiangyu Zhang, and Jian Sun. CrowdHuman: A benchmark for detecting human in a crowd. *arXiv preprint arXiv:1805.00123*, 2018.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? Investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 6048–6058, 2023.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *Transactions on Machine Learning Research (TMLR)*, 2022.

Jiazheng Xu, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. ImageReward: Learning and evaluating human preferences for text-to-image generation. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023.

Shuo Yang, Ping Luo, Chen-Change Loy, and Xiaoou Tang. WIDER FACE: A face detection benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5525–5533, 2016.

A Full Experimental Specification

Main grid. Template: *"A photograph of exactly* $\{N\}$ *people* $\{a\}$ $\{s\}$, $\{l\}$*, facing the camera, no other people in the scene, realistic, highly detailed"*, with $N \in \{1, 2, 4, 8, 16, 32, 64\}$; scenes $s \in$ {on a city street, on a sandy beach, in a modern open-plan office}; lighting $l \in$ {in bright daylight, at night under artificial lighting}; activities $a \in$ {walking, standing and talking}; seeds $\{0, \ldots, 4\}$.

Identity probe. *"A tightly framed group photo of exactly* $\{N\}$ *people standing close together,* $\{l\}$*, facing the camera, realistic, highly detailed"*, $N \in \{2, \ldots, 64\}$, $l \in$ {studio lighting, outdoor daylight}, seeds $\{0, \ldots, 9\}$.

Cars grid. *"A photograph of exactly* $\{N\}$ *cars parked* $\{s\}$, $\{l\}$*, no other cars in the scene, realistic, highly detailed"*, $s \in$ {along a city street, in a large open parking lot}, same N and lighting, seeds $\{0, \ldots, 4\}$.

Negative prompt for all conditions: *"cartoon, painting, illustration, low quality, blurry"*. Generation: SDXL base 1.0 (1024^2, 30 steps, CFG 7.0), SD 1.5 (512^2, 50 steps, CFG 7.5), PixArt-XL-2-1024-MS (1024^2, 20 steps, CFG 4.5). Detection: YOLO11x at 1536px inference, confidence 0.2, max $1{,}000$ detections; SCRFD (buffalo_l) at 1024^2 detection resolution, confidence 0.5; ArcFace embeddings restricted to structurally valid faces with min side $\geq$ 32px. Blur-validity thresholds are calibrated per size bucket on real faces (Table 3).

B Complete Scaling and Onset Fits

Table 4 reports both candidate functional forms for every recall and validity metric, on the main and cars grids; Table 5 reports every segmented-regression onset fit, including non-significant ones.

Table 3: Size-conditional blur-validity thresholds: 5th percentile of Laplacian variance over 4,000 real WIDER FACE crops per size bucket (crops resized to 112^2).

Face size bucket	threshold
$\geq$ 8px	1.8
$\geq$ 16px	2.2
$\geq$ 24px	2.5
$\geq$ 32px	3.3
$\geq$ 48px	5.0
$\geq$ 64px	9.3
$\geq$ 96px	27.3
$\geq$ 128px	80.4

Table 4: Complete scaling fits: both functional forms for every recall/validity metric (main and cars grids). The final column gives the AIC-selected form and the bootstrap 95% CI of its decay parameter.

Model	Metric	Power law $cN^{-\alpha}$	Exponential $ce^{-\lambda N}$	Selected [CI]
SDXL	person recall	$\alpha = 0.148$, AIC -547	$\lambda = 0.0114$, AIC -573	exp. $[0.009, 0.014]$
SDXL	face recall	$\alpha = 0.375$, AIC -188	$\lambda = 0.0258$, AIC -239	exp. $[0.021, 0.031]$
SDXL	face validity	$\alpha = 0.043$, AIC -808	$\lambda = 0.0031$, AIC -814	exp. $[0.002, 0.005]$
SDXL	face recall corrected	$\alpha = 0.261$, AIC -194	$\lambda = 0.0194$, AIC -235	exp. $[0.014, 0.024]$
SD 1.5	person recall	$\alpha = 0.321$, AIC -477	$\lambda = 0.0209$, AIC -482	exp. $[0.018, 0.024]$
SD 1.5	face recall	$\alpha = 0.616$, AIC -252	$\lambda = 0.0411$, AIC -291	exp. $[0.037, 0.046]$
SD 1.5	face validity	$\alpha = 0.022$, AIC -731	$\lambda = 0.0011$, AIC -730	pow. $[-0.001, 0.044]$
SD 1.5	face recall corrected	$\alpha = 0.506$, AIC -258	$\lambda = 0.0344$, AIC -301	exp. $[0.030, 0.039]$
PixArt-α	person recall	$\alpha = 0.342$, AIC -522	$\lambda = 0.0235$, AIC -563	exp. $[0.022, 0.026]$
PixArt-α	face recall	$\alpha = 0.584$, AIC -201	$\lambda = 0.0345$, AIC -215	exp. $[0.030, 0.038]$
PixArt-α	face validity	$\alpha = 0.016$, AIC -483	$\lambda = -0.0001$, AIC -482	pow. $[-0.022, 0.057]$
PixArt-α	face recall corrected	$\alpha = 0.469$, AIC -202	$\lambda = 0.0282$, AIC -220	exp. $[0.025, 0.032]$
SDXL (cars)	object recall	$\alpha = 0.208$, AIC -60	$\lambda = 0.0177$, AIC -73	exp. $[0.012, 0.023]$
SD 1.5 (cars)	object recall	$\alpha = 0.561$, AIC 14	$\lambda = 0.0441$, AIC 20	pow. $[0.350, 0.778]$
PixArt-α (cars)	object recall	$\alpha = 0.406$, AIC -133	$\lambda = 0.0266$, AIC -138	exp. $[0.022, 0.031]$

C Detector Calibration Detail

Tables 6 and 7 give the full per-bin controls (with sample counts) summarized in the main text: face detection degrades substantially with density on real images, while car detection does not. Table 8 illustrates why detector *configuration* is itself a confound: default settings miss 85% of the people the same detector finds at dense-crowd settings in the same image. All measurements in this paper — on generated and real images alike — use the dense-crowd configuration.

D Threshold Sensitivity of Duplicate Emergence

Figure 8 shows the image-level duplicate-presence rate across ArcFace thresholds $\tau \in \{0.3, 0.4, 0.5, 0.6\}$. The rise with N is present at every threshold; τ shifts the level, not the shape.

E Attention Probe Detail

We instrument all cross-attention (`attn2`) modules of the SDXL UNet, computing explicit attention probabilities for the counting-noun token ("person"/"people"), averaged over heads and layers within each spatial resolution, accumulated over the second half of denoising steps (early steps are near-uniform for all tokens). Self-attention is left untouched (sliced processors). Per image we record the map at the mid-block resolution (most semantically structured) and the highest captured resolution; person regions come from the same YOLO configuration as the main pipeline. Table 9 gives the full metrics; Figure 9 visualizes the maps.

Table 5: All segmented-regression onset fits, including non-significant ones. "Significant" means the two-segment model beats a single line by AIC margin > 2.

Model	Statistic	N^*	95% CI	significant
SDXL	mean max sim	4.7	$[2.4, 17.1]$	yes
SD 1.5	mean max sim	12.9	$[2.4, 41.3]$	yes
PixArt-α	mean max sim	6.2	$[4.0, 8.0]$	yes
PixArt-α	max pair sim	6.8	$[2.7, 8.0]$	yes
PixArt-α	has dup@0.4	8.0	$[6.5, 41.3]$	yes
SDXL (probe)	mean max sim	4.5	$[2.4, 44.8]$	no
SDXL (probe)	max pair sim	11.0	$[2.4, 47.2]$	no
SDXL (probe)	has dup@0.4	21.9	$[4.5, 45.9]$	yes
SD 1.5 (probe)	mean max sim	3.8	$[2.4, 13.0]$	no
SD 1.5 (probe)	max pair sim	8.0	$[2.4, 15.9]$	no
SD 1.5 (probe)	has dup@0.4	16.0	$[2.4, 33.4]$	no
PixArt-α (probe)	mean max sim	12.2	$[2.4, 39.2]$	no
PixArt-α (probe)	max pair sim	10.5	$[3.1, 17.8]$	yes
PixArt-α (probe)	has dup@0.4	19.7	$[2.4, 43.6]$	no

Table 6: WIDER FACE detector-degradation control, full detail (dense-crowd settings, 1024px detection resolution).

Density bin	Count ratio	Matched recall	images
1	0.960	0.880	300
2	0.980	0.937	300
4	0.882	0.859	300
8	0.823	0.800	300
16	0.748	0.719	300
32	0.712	0.683	175
64	0.621	0.588	99

F Birthday-Collision Model

If each of an image's n_e embeddable faces draws an identity independently and uniformly from an effective pool of size K, the probability that at least one pair shares an identity is $P(\text{dup}) = 1 - \prod_{i=1}^{n_e-1}(1 - i/K) \approx 1 - e^{-n_e(n_e-1)/2K}$. We fit K by maximizing the Bernoulli likelihood of per-image duplicate presence (at τ=0.4) over each image's observed n_e, and compare against a two-parameter logistic in $\log_2 N$ by AIC (Section 5.3). Figure 10 overlays the fits on the empirical rates.

G Additional Qualitative Results

Figure 11 shows the highest-similarity within-image identity pairs per model; Figures 12 and 13 extend the main-text face-crop grid (Figure 4) to SD 1.5 and PixArt-α.

H Design Evolution: What the Pilots Caught

Constructing a valid density benchmark required several iterations; we document the failure modes so future benchmarks can avoid them. (1) *Scene wording controls implied population*: a "busy city street" scene produced $\sim$30 background pedestrians on an N=1 prompt (Figure 14a), motivating neutral scene wording plus a uniform "no other people" clause. (2) *Lone subjects face away*: without a camera-facing clause, only 1 of 12 pilot N=1 images contained a detectable face (Figure 14b). (3) *Blur validity conflates size with quality*: a flat sharpness threshold rejected 94% of detected faces; thresholds must be calibrated per face size on real images (Table 3). (4) *Detector defaults silently fail in crowds* (Table 8). (5) *Identity analysis needs a resolution floor*: main-grid faces at $N \geq 32$ fall below ArcFace's useful scale, motivating the close-up identity probe.

Table 7: COCO val2017 car-detection control: no density degradation for cars (bins ≥ 24 true cars do not occur in COCO val).

Density bin	Count ratio	Matched recall	images
1	1.019	0.631	160
2	1.074	0.748	178
4	1.085	0.733	121
8	1.013	0.777	50
16	1.417	0.817	26

Table 8: Detector-settings sensitivity: persons found by YOLO11x in a single pilot $N{=}64$ generated image containing $\sim$100 visible people.

YOLO configuration	persons detected
640px, conf 0.30, max 300 (defaults)	15
1024px, conf 0.30, max 300	31
1024px, conf 0.15, max 1000	63
1536px, conf 0.15, max 1000	98

I Factor Analysis

Table 10 decomposes recall variance over the full grid. Scene modulates recall levels for the 1024^2 models; density is the largest factor for SD 1.5 and, crucially, the only factor whose effect is a universal monotone trend: mean person recall declines from $N{=}4$ to $N{=}64$ in all 36 scene$\times$lighting$\times$activity cells across the three models (median decline ratio 0.32–0.55, maximum 0.82).

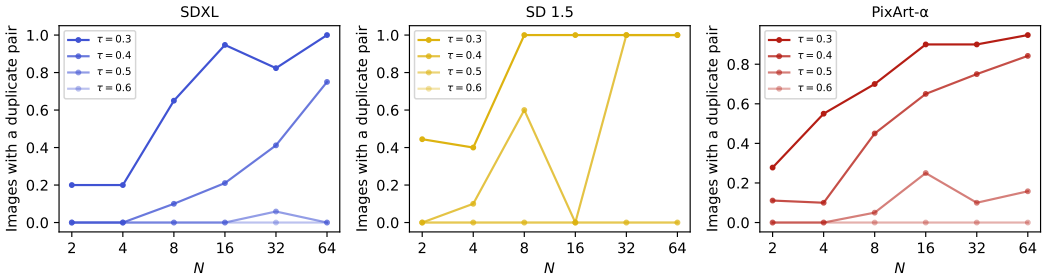


Figure 8: Duplicate-presence rate vs. N across similarity thresholds (identity probe).

Table 9: Attention-probe metrics (SDXL, counting-noun token, mid-block cross-attention, second half of denoising; uniform-map entropy is 6.931). Attention mass in person regions tracks person pixel area at every N.

N	det. persons	entropy	mass in persons	mass/person	person area frac.
1	10.0	6.9302	0.188	0.0267	0.182
2	4.7	6.9305	0.173	0.0432	0.153
4	9.3	6.9296	0.359	0.0450	0.337
8	11.0	6.9302	0.443	0.0407	0.429
16	40.3	6.9305	0.459	0.0129	0.443
32	66.0	6.9303	0.406	0.0061	0.399
64	65.0	6.9305	0.387	0.0060	0.382

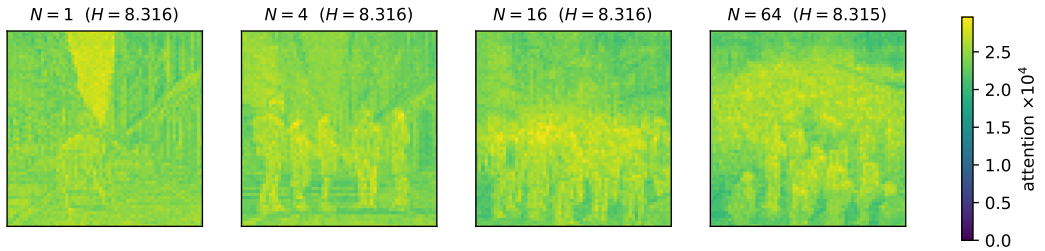


Figure 9: Counting-token cross-attention maps (SDXL, seed 0) at increasing N. H is spatial entropy; the uniform-map value is 6.931. The conditioning signal is spatially undifferentiated at every density.

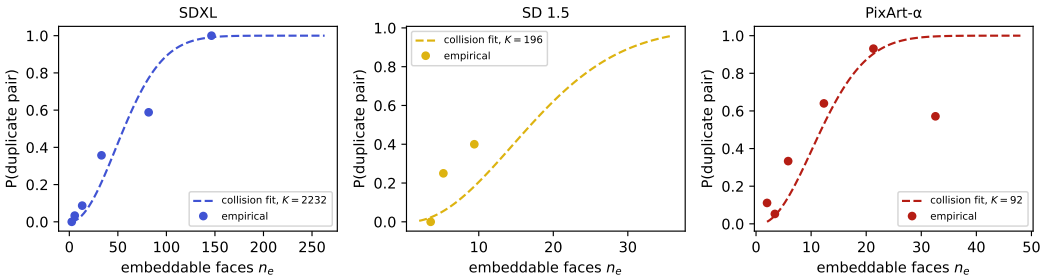


Figure 10: Birthday-collision fits (dashed) vs. empirical duplicate-presence rates binned by embeddable-face count.

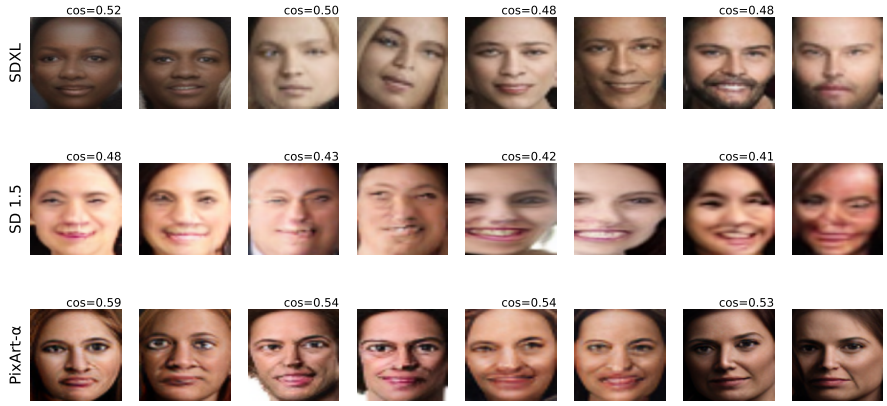


Figure 11: Duplicate-identity gallery: for each model, the four highest-similarity ArcFace pairs found *within single generated images* of the identity probe. These are distinct requested entities rendered with near-identical identities.

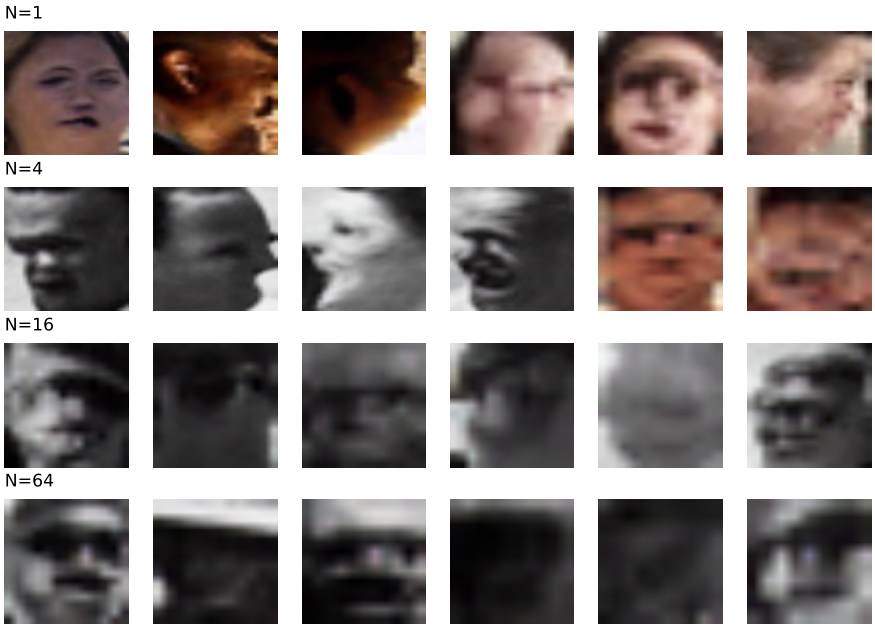


Figure 12: Detected face crops at increasing density, SD 1.5 (seed 0).

Table 10: Variance decomposition (η^2) of recall over the full main grid.

Model	Metric	N	scene	lighting	activity	seed
SDXL	person recall	0.102	0.131	0.007	0.005	0.011
SDXL	face recall	0.121	0.143	0.001	0.084	0.007
SD 1.5	person recall	0.245	0.044	0.016	0.037	0.030
SD 1.5	face recall	0.319	0.033	0.032	0.059	0.033
PixArt-α	person recall	0.212	0.088	0.021	0.000	0.005
PixArt-α	face recall	0.107	0.126	0.053	0.069	0.004

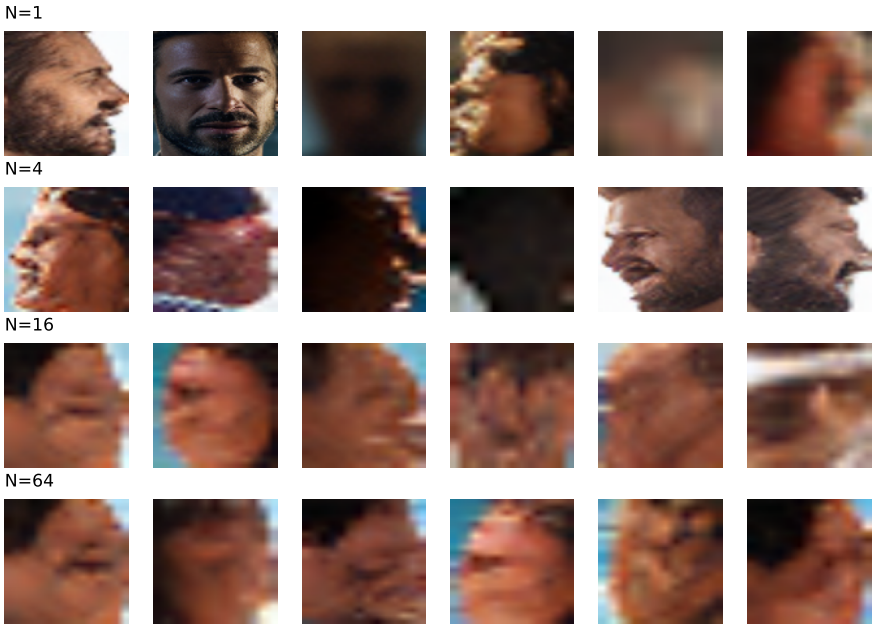


Figure 13: Detected face crops at increasing density, PixArt-α (seed 0).



Figure 14: Pilot confounds, reproduced with the pilot templates (SDXL, seed 0). (a) Scene wording alone induces a crowd at $N{=}1$. (b) With counts controlled but no facing clause, the lone subject renders from behind.