BOLDNESS: EVALUATING MODELS UNDER LABEL-CHANGING EDITS VIA DIFFUSION LATENT OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

We define *boldness* as a model's ability to predict the new oracle label under valid nearby label-changing perturbations, complementing robustness under label-preserving perturbations. Evaluating boldness requires constructing natural, source-faithful samples whose oracle labels change. We characterize these requirements and formulate boldness-sample construction as a constrained optimization problem. To make this construction practical, we develop a plug-in diffusion method that optimizes the latent seed of a fixed pretrained conditional generator. The method combines target-class conditioning with original-label loss to search for label-changing edits on which the model retains the source label, while regularizing the latent seed around a source-inversion solution to promote source faithfulness. Target semantics and source faithfulness are validated before evaluation. Experiments on image classification and object detection reveal substantial boldness failures across the evaluated model families, including robustness-enhanced models, highlighting the need to evaluate boldness alongside clean accuracy and robustness.

1 INTRODUCTION

We consider a scenario in which autonomous driving systems and human drivers both use speed-limit signs to regulate speed, and call the label determined by human judgment the *oracle label*. Assume that a model correctly predicts the oracle label of the original sign. As shown in Figure 1, altering the sign can produce two failure modes: **(A1)** a *label-preserving perturbation* leaves the oracle label unchanged but causes the model to predict a different label; **(A2)** a *label-changing perturbation* changes the oracle label, but the model fails to predict the new label. We call the ability to predict the new oracle label after a valid nearby label change *boldness*.

Correctness is routinely evaluated, and robustness has been studied extensively (Goodfellow et al., 2015; Madry et al., 2018; Croce et al., 2020). By contrast, boldness has received comparatively limited attention (Jacobsen et al., 2019; Tramèr et al., 2020; Borji, 2022a). Closely related work searches for substantial input changes that preserve the source prediction (Borji, 2022b; Nie et al., 2025; Borji, 2026). However, a distance lower bound can admit an unrelated image from another class that the model misclassifies as the original label. Such an image can satisfy the large-distance condition without being a meaningful edit of the particular input.

This motivates specifying the requirements of a boldness failure before constructing the attack. In Sec. 2, we distinguish boldness from correctness and robustness and identify four requirements: label-change validity, image validity, source faithfulness, and model failure. We then formulate boldness-sample construction as a constrained optimization problem and make the formulation prac-



Figure 1: **Failures under label-preserving and label-changing perturbations.** (A1) A slight rotation preserves the oracle label but changes the model prediction. (A2) A stronger transformation changes the oracle label from "80" to "08", but the model retains the source prediction.

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.

tical using a fixed pretrained conditional image generator. The resulting method uses target conditioning, source inversion, and latent optimization to construct source–target transitions and search for conservative boldness failures. We further extend boldness evaluation to object detection by constructing object-level boldness samples within annotated image regions. Experiments on image classification and object detection reveal substantial boldness failures across model families and robustness-enhanced models. Our contributions are:

- **We provide an explicit formalization of boldness.** We define boldness as a model's ability to predict the new oracle label under valid nearby label-changing perturbations, distinguishing it from correctness and robustness. Unlike prior heuristic or large-distance constructions that do not enforce source faithfulness, our formulation makes explicit that a boldness failure must occur on a natural, label-changing sample that remains meaningfully linked to its source.
- **We develop a practical constrained formulation and latent optimization algorithm for boldness evaluation.** The formulation jointly captures label-change validity, image validity, source faithfulness, and the evaluated-model failure condition. The resulting algorithm uses a fixed pretrained conditional generator and latent optimization to construct source–target transitions and search for boldness failures.
- **We uncover severe failures to adapt predictions after semantic changes.** Across image classification and object detection, we observe boldness failures across model families and robustness-enhanced models, showing that high clean accuracy or robustness does not guarantee boldness. In object detection, high box retention together with low boldness mAP further indicates that the failure primarily lies in updating the predicted category rather than localizing the edited object.

2 Problem Definition of Boldness

2.1 Definitions of Correctness, Robustness, and Boldness

To formalize the distinction between model behavior under label-preserving and label-changing perturbations, let $\mathcal{X}$ denote the input space, $\mathcal{Y}$ the label space, and $o : \mathcal{X} \to \mathcal{Y}$ the oracle labeling function. Let $f : \mathcal{X} \to \mathbb{R}^K$ denote the model logits, and let $\hat{y}(\mathbf{x}) = \arg\max_k f_k(\mathbf{x})$ denote the predicted label. We call an image valid if it is a natural image with a clear oracle label. For a valid original sample $\mathbf{x}$, write $y = o(\mathbf{x})$. Let $\mathcal{N}_R(\mathbf{x})$ denote the set of valid nearby samples whose oracle label remains y, and let $\mathcal{N}_B(\mathbf{x})$ denote the set of valid nearby samples whose oracle label differs from y. We consider three model properties:

(P0) *Correctness.* The model predicts the correct oracle label for the original input: $\hat{y}(\mathbf{x}) = o(\mathbf{x}) = y$.

(P1) *Robustness.* For every valid nearby sample whose oracle label remains y, the model should also predict it as y: $\forall \mathbf{x}_R \in \mathcal{N}_R(\mathbf{x}), \hat{y}(\mathbf{x}_R) = y$.

(P2) *Boldness.* For every valid nearby sample whose oracle label changes, the model should predict its new label: $\forall \mathbf{x}_B \in \mathcal{N}_B(\mathbf{x}), \hat{y}(\mathbf{x}_B) = o(\mathbf{x}_B)$.

Correspondingly, a *robustness attack* searches for a *robustness sample* $\mathbf{x}_R \in \mathcal{N}_R(\mathbf{x})$ on which $\hat{y}(\mathbf{x}_R) \neq y$, whereas a *boldness attack* searches for a *boldness sample* $\mathbf{x}_B \in \mathcal{N}_B(\mathbf{x})$ on which $\hat{y}(\mathbf{x}_B) \neq o(\mathbf{x}_B)$. As shown in Figure 2, the two attacks probe different model behaviors, and robustness and boldness failures can coexist.

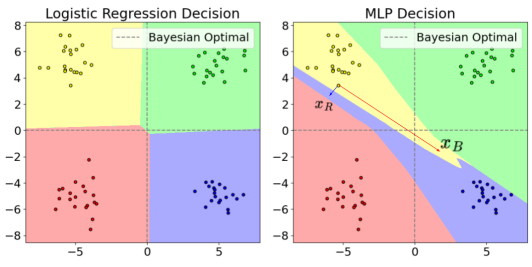


Figure 2: **Robustness and boldness failures can coexist.** Left: the logistic-regression boundary matches the oracle boundary, so predictions remain stable under label-preserving changes and adapt under label-changing changes. Right: the multilayer perceptron exhibits both failures, changing along a label-preserving direction while retaining the original label after a label-changing perturbation.

Relation to prior large-distance constructions. Prior work closest to boldness studies a reversal of the usual adversarial robustness objective: rather than searching within a small perturbation budget for an input that changes the prediction, it searches for a substantially changed input on which the

original prediction is preserved (Borji, 2022b; Nie et al., 2025; Borji, 2026). In particular, Nie et al. (2025) makes this reversal explicit by minimizing the original-label loss subject to a lower bound on input distance. The lower bound enforces a large input change, but does not require the resulting image to be a meaningful semantic modification of $\mathbf{x}$. Indeed, consider an unrelated natural image $\tilde{\mathbf{x}}$ from another class such that $o(\tilde{\mathbf{x}}) \neq y$, $\hat{y}(\tilde{\mathbf{x}}) = y$, and ℓ_p distance $\|\tilde{\mathbf{x}} - \mathbf{x}\|_p$ exceeds the prescribed lower bound. Such an image satisfies the label-change, original-prediction, and distance conditions, even though it is not a meaningful modification of $\mathbf{x}$. A classifier $\hat{y}$ with more such cross-class errors offers more opportunities for this trivial construction. Thus, large pixel-space distance does not characterize the source-specific nearby label-changing samples required for boldness evaluation.

2.2 Requirements and Challenges for Constructing Boldness Samples

This motivates defining the construction from the requirements of a boldness failure rather than from perturbation magnitude alone. Constructing a boldness sample involves three hard requirements and one soft requirement. First, **label-change validity** is a hard requirement that the sample have an oracle label different from y. Second, **image validity** is a hard requirement that the sample remain a natural image. Third, the **boldness-failure condition** is a hard requirement that the evaluated model violate (P2); in this work, we focus on the conservative case in which the model continues to predict the original label y. Finally, **source faithfulness** requires the sample to remain meaningfully linked to $\mathbf{x}$, operationalizing the "nearby" relation in $\mathcal{N}_B(\mathbf{x})$. Because there is no universal semantic radius that separates a faithful edit from an unrelated sample, we encourage source faithfulness through a soft objective rather than impose a fixed semantic radius. Constructing a boldness sample that satisfies all four requirements raises two main challenges.

Challenge 1: Satisfying heterogeneous requirements. These requirements concern different aspects of the edit: its oracle label, naturalness, relation to the source, and model prediction. They also create tension: changing the oracle label requires modifying task-relevant content, whereas source faithfulness limits unrelated changes, and the failure condition requires the model to retain the source prediction despite the semantic change. A single perturbation magnitude cannot capture these demands.

Challenge 2: Making the construction computable. The requirements above do not directly specify a computable construction. The natural-image set and oracle labels are unavailable during optimization, and source faithfulness lacks a canonical computable metric.

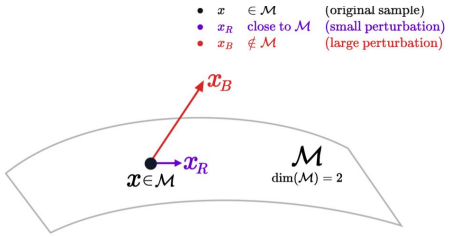


Figure 3: **Large perturbations may leave the natural-image manifold.** Natural images occupy a structured subset $\mathcal{M} \subseteq \mathcal{X}$. Small perturbations may remain on or near this set, whereas large perturbations may leave it and produce unnatural images.

3 Our Method

We now turn the requirements in Sec. 2.2 into a computable search for boldness failures. To address Challenge 1, we combine the three hard requirements in a constrained formulation and minimize source distance over its feasible set. To address Challenge 2, we use a fixed conditional generator to promote image validity, target conditioning to direct the semantic change, and source inversion to define a source-relative latent penalty.

3.1 Conceptual Boldness Formulation

Given an original image $\mathbf{x}$ with oracle label $y = o(\mathbf{x})$, let $\mathcal{M} \subseteq \mathcal{X}$ denote the natural-image manifold and $D(\mathbf{x}, \mathbf{x}_B)$ a measure of image deviation to be instantiated below. We first leave the changed oracle label unspecified and search for the conservative boldness failure described in Sec. 2.2:

$$\min_{\mathbf{x}_B} D(\mathbf{x}, \mathbf{x}_B) \quad \text{s.t. } \mathbf{x}_B \in \mathcal{M}, \quad o(\mathbf{x}_B) \neq y, \quad \hat{y}(\mathbf{x}_B) = y. \tag{1}$$

By contrast, the large-distance objective of Nie et al. (2025) described in Sec. 2.1 has the form

$$\min_{\mathbf{x}_B} \ell(f(\mathbf{x}_B), y) \quad \text{s.t. } \|\mathbf{x}_B - \mathbf{x}\|_p \geq \delta, \tag{2}$$

where $\delta > 0$ is a prescribed lower bound. The main difference is the role of distance: we minimize deviation from the particular input subject to explicit image-validity, label-change, and prediction constraints, rather than require a minimum perturbation size. This matches the purpose of a boldness attack: an unrelated misclassified image reveals an existing classification error, whereas a meaningful edit tests the model's response to a change of the particular input. Minimizing D keeps this relation to $\mathbf{x}$ explicit in the search, rather than accepting a candidate merely because it is far away.

The condition $o(\mathbf{x}_B) \neq y$ specifies which label to leave, but not which alternative the generator should produce. We make the desired change explicit by considering target-conditioned subproblems indexed by $y' \in \mathcal{Y} \setminus \{y\}$. Let $\mathcal{T}(\mathbf{x}) \subseteq \mathcal{Y} \setminus \{y\}$ denote the target labels selected for search. For each fixed $y' \in \mathcal{T}(\mathbf{x})$, we solve

$$\min_{\mathbf{x}_B} D(\mathbf{x}, \mathbf{x}_B) \quad \text{s.t. } \mathbf{x}_B \in \mathcal{M}, \quad o(\mathbf{x}_B) = y', \quad \hat{y}(\mathbf{x}_B) = y. \tag{3}$$

Restricting $\mathcal{T}(\mathbf{x})$ can miss failures for unsearched labels; it limits search coverage rather than establishing (P2) on those labels.

3.2 Boldness Evaluation via Constrained Plug-in Optimization

The conceptual formulation in Eq. (3) specifies what a boldness sample should satisfy, but its natural-image manifold, oracle label, and image distance are not directly available during optimization. We therefore instantiate these components progressively using a fixed pretrained conditional generator, with each step yielding a more computable version of the preceding formulation.

Image validity via pretrained text-to-image models. Large-scale text-to-image models, such as Stable Diffusion (Rombach et al., 2022) and SDXL (Stability AI, 2023), are trained on large and diverse collections of natural images to transform samples from a simple input prior into samples from a learned image distribution. Having been trained on broad image distributions, such models provide a reusable generative prior over diverse natural visual content, making them a practical mechanism for promoting image validity. Let $G_\theta : \mathcal{Z} \to \mathcal{X}$ denote this fixed generative map (with its conditioning input temporarily suppressed). We use its range as a practical approximation to $\mathcal{M}$: a candidate in this range is represented as $\mathbf{x}_B = G_\theta(z)$ for some seed $z \in \mathcal{Z}$. Because the generator is trained using seeds sampled from its input prior, its behavior is best supported in regions that are typical under that prior. For the standard Gaussian seed $z \sim \mathcal{N}(0, I_m)$, concentration of measure implies that $\|z\|_2$ lies sharply around $\sqrt{m}$. Unconstrained optimization can move z away from this high-probability region and query the fixed generator where its behavior is less constrained by training. Following Wan et al. (2025), we therefore restrict the optimized seed to $\mathcal{S}_\epsilon := \{z : (1-\epsilon)\sqrt{m} \leq \|z\|_2 \leq (1+\epsilon)\sqrt{m}\}$. Restricting the search to generator outputs produced from typical prior seeds provides a practical instantiation of the image-validity requirement $\mathbf{x}_B \in \mathcal{M}$. The conceptual problem becomes

$$\min_{z \in \mathcal{S}_\epsilon} D(\mathbf{x}, G_\theta(z)) \quad \text{s.t.} \quad o(G_\theta(z)) = y', \quad \hat{y}(G_\theta(z)) = y. \tag{4}$$

Label-change validity via target text conditioning. After restricting the search to plausible generator outputs, we still need a computable mechanism for directing those outputs toward the specified target label y'. Instead of querying the inaccessible oracle during optimization, we encode the requested oracle label in the generator's condition. Modern foundation image generators commonly use text conditioning to control the semantic content of generated images across a large vocabulary of visual concepts. We therefore make the conditioning input explicit and write the generator as $G_\theta(z; c)$. For a specified target label $y' \neq y$, we construct a text condition $c_{y'}$ from the target class name and keep it fixed throughout optimization: $\mathbf{x}_B = G_\theta(z; c_{y'})$. Target conditioning thus provides a computable surrogate for the oracle-label constraint: it directs candidate generation toward y' without requiring queries to o during optimization. The optimization problem therefore becomes

$$\min_{z \in \mathcal{S}_\epsilon} D(\mathbf{x}, G_\theta(z; c_{y'})) \quad \text{s.t.} \quad \hat{y}(G_\theta(z; c_{y'})) = y. \tag{5}$$

Source faithfulness via inversion and seed-space distance. Target conditioning specifies which semantics should change, but it does not by itself ensure that the resulting target-class image remains tied to the particular source $\mathbf{x}$. We next instantiate the source-distance term. Pixel-space distance is poorly aligned with source faithfulness because visually irrelevant low-level changes can produce large pixel differences (Zhang et al., 2018). This motivates learned perceptual distances, which compare images in feature spaces that suppress some low-level variation while retaining perceptually meaningful structure (Zhang et al., 2018; Laidlaw et al., 2021). Following the same principle, we instead measure source deviation in a representation space associated with the generator.

Specifically, we invert the original image under its original condition c_y to obtain a seed z^* satisfying $G_\theta(z^*; c_y) \approx \mathbf{x}$. This provides a source-specific reference in the same seed space in which z is optimized. Keeping z close to z^* therefore discourages the optimization from drifting to an unrelated target-class generation. Although distance could also be defined in downstream generator representations, we use the seed space because it is computationally simple for optimizing z. We implement $D(\mathbf{x}, \mathbf{x}_B)$ using the seed-space distance $\|z - z^*\|_2^2/m$, where m is the seed dimension. The optimization problem becomes

$$\min_{z \in \mathcal{S}_\epsilon} \ \|z - z^*\|_2^2/m \quad \text{s.t.} \quad \hat{y}(G_\theta(z; c_{y'})) = y. \tag{6}$$

Boldness failure via a classification surrogate loss. The remaining requirement is to search for the conservative boldness failure $\hat{y}(\mathbf{x}_B) = y$, in which the evaluated model retains the original label after the image semantics have changed toward y'. Cross-entropy $\ell(f(\mathbf{x}_B), y)$ provides a differentiable classification surrogate for this discrete prediction constraint. Composing it with the generator yields $\ell(f(G_\theta(z; c_{y'})), y)$. The target condition and evaluated-model loss therefore serve different and opposing roles: $c_{y'}$ directs the generated image toward the target semantics, whereas the source-label loss searches within those edits for cases in which the evaluated model remains attached to y. Combining this loss with the source-distance term yields our final plug-in objective

$$\boxed{\min_{z \in \mathcal{S}_\epsilon} \ \ell\left(f(G_\theta(z; c_{y'})), y\right) + \lambda_z \|z - z^*\|_2^2/m} \tag{7}$$

where $\lambda_z \geq 0$ controls source regularization. We solve Eq. (7) using projected gradient descent initialized at $z^{(0)} = \Pi_{\mathcal{S}_\epsilon}(z^*)$, projecting each update back onto $\mathcal{S}_\epsilon$.

Exhaustively solving Eq. (7) for every $y' \neq y$ can strengthen the search, but a single valid failure already constitutes a counterexample to (P2). For classification, we use $\mathcal{T}(\mathbf{x}) = \{y'\}$, where $y' = \arg\max_{k \in \mathcal{Y} \setminus \{y\}} f_k(\mathbf{x})$. This chooses the highest-logit alternative to the original label; when the original image is classified correctly, it is the class with the second-largest logit. It requires only one target-conditioned optimization per image. The low same-model boldness accuracy in Tab. 4 shows that this restricted search already exposes substantial failures. Searching more labels may reveal additional failures; the reported accuracy characterizes the selected targets and retained edits, not an exhaustive evaluation of (P2).

The generator parameterization, target conditioning, seed-space distance, and evaluated-model loss are practical computational instantiations of the conceptual requirements rather than exact enforcement of them. In particular, target conditioning and the generative prior do not by themselves certify target semantics, source faithfulness, or image validity. We therefore verify these properties for each generated candidate before including it in the boldness evaluation, as described in Sec. C.

3.3 Boldness for Object Detection

Boldness is not specific to single-label classification, so we also consider object detection, where each prediction contains both a semantic category and a spatial localization, typically produced by classification and bounding-box regression. A general notion of detection boldness could consider oracle changes to both components. In this work, we restrict attention to category boldness: the oracle category changes from y_i to y_i', while the object is intended to remain in approximately the same spatial region. Accordingly, a bold detector should continue to localize the edited object while updating its predicted category to y_i'; retaining y_i on a spatially matched detection constitutes the conservative category-boldness failure studied here.

Let $\{(b_i, y_i)\}_{i=1}^M$ denote the annotated objects in an image, where b_i is a bounding box and y_i is its original class. For an eligible object i, we choose a candidate target category $y_i' \neq y_i$ and construct an edited image $\mathbf{x}_B^{(i)}$ by restricting the semantic edit to b_i, while preserving pixels outside the edited region. The detector is then evaluated on the edited image. A successful response requires a detection that spatially matches b_i and predicts category y_i'.

4 Empirical Evaluation

In this section, we empirically evaluate boldness along three dimensions: its prevalence across standard image classifiers, its relationship with robustness, and its extension to object detection. Classification evaluates adaptation to changes in image-level labels, whereas object detection tests whether

localized category predictions adapt to semantic edits in multi-object scenes. We first present classification experiments and then extend the evaluation to verified local edits in object detection.

4.1 CLASSIFICATION

Datasets. For image classification, we use the ImageNet-100 subset introduced by Tian et al. (2020), which contains 100 classes selected from ImageNet-1K. From these classes, we sample 100 original images per category, yielding 10,000 original images in total. For all classifiers, prediction and target selection are restricted to the same 100 ImageNet-100 categories. For ImageNet-1K pretrained ResNet-50 and ViT-B/16, we retain the logits corresponding to these categories. For CLIP, we construct a zero-shot classifier over the same 100 class names. For each evaluated model, both source and edited images are processed using the preprocessing prescribed by its released checkpoint.

Evaluated Models. For standard classification, we adopt and evaluate ResNet-50 and ViT-B/16 using their publicly released TorchVision checkpoints (He et al., 2016; Dosovitskiy et al., 2020), and CLIP ViT-L/14 using the official OpenAI checkpoint (Radford et al., 2021) and follow the zero-shot ImageNet classification protocol. These models represent convolutional, transformer-based, and vision–language pretrained architectures, respectively. To examine boldness performance of robust models, we additionally evaluate several ResNet models with two robustness-enhanced approaches. We use the ℓ_∞-robust ($\epsilon = 8/255$) and ℓ_2-robust ($\epsilon = 3.0$) ResNet-50 to represent adversarial robust models, with their checkpoints taken from the publicly released MadryLab robustness model zoo (Engstrom et al., 2019). We also use the AugMix ResNet-50 to represent natural robust model, with its checkpoint taken from the official release (Hendrycks et al., 2020).

Edit Generation and Optimization Settings. For the fixed generative model, we use Stable Diffusion 2.1 based on the latent diffusion framework (Rombach et al., 2022). Following the construction in Sec. 3.1, original and target class labels are used as the text conditions c_y and $c_{y'}$. We use 5 inversion steps to obtain the source-associated latent z^*. We solve Eq. (7) using 50 projected-gradient updates with step size $\eta = 0.01$ and source-regularization coefficient $\lambda_z = 0.003$. We set the Gaussian-shell width to $\epsilon = 0.025$, following FMPlug (Wan et al., 2025), and keep this value fixed throughout all experiments. For each original image, we choose y' as the highest-scoring non-original class of the model used for edit optimization.

Evaluation Metrics. Tab. 1 summarizes the metrics. Only candidates satisfying all three thresholds in Tab. 2 are retained. The validity pass rate is the fraction of generated candidates that are retained; BA is evaluated only on retained edits.

Table 1: Classification evaluation metrics.

Metric	Definition and scope
FID ↓	Feature-distribution discrepancy from target-class images (Heusel et al., 2017).
TIFA ↑	Consistency with the target description (Hu et al., 2023).
DreamSim ↑	Perceptual similarity to the original image (Fu et al., 2023).
Edit-vs-random gap ↑	Source similarity minus the mean for random target generations; see Sec. C.
BA ↑	$\mathrm{BA}_{\mathrm{cls}} = \Pr[\hat{y}(\mathbf{x}_B) = y']$ over retained edits with verified target labels y'.

Metric	Mean ↑	Std	Threshold
TIFA target score	0.92	0.09	≥ 0.8
DreamSim source similarity	0.61	0.12	≥ 0.5
Edit-vs-random gap	0.21	0.08	> 0

Table 2: Per-image validity of generated ImageNet edits. Only candidates satisfying all three criteria are retained for boldness evaluation.

Experimental Results for Classification. We assess edit quality before prediction accuracy.

(1). Our edits better match the clean-image distribution. Tab. 3 compares FID for our validated semantic edits and the pixel-space ℓ_∞ baseline, with a clean–clean reference. Our edits have a lower FID than the baseline, indicating closer agreement with the reference image distribution under this metric; **(2). Most generated edits satisfy the validity protocol.** Tab. 2 summarizes the per-image validation statistics, while Fig. 4 shows representative retained edits. Under this protocol, 90.4% of



(a) espresso→bowl (b) conker→hair slide (c) fire truck→mob. home (d) toyshop→grocery

Figure 4: **Representative validated ImageNet edits.** Each edit acquires the verified target semantics while remaining source-related, yet the model retains the source prediction.

candidates are retained for ResNet50 model. The remaining 9.6% fail at least one check and are not admitted as valid label-changing edits. We exclude their predictions from BA to avoid conflating uncertainty about edit validity with model errors on valid semantic changes.

(3). Boldness failures occur across classifier families and transfer across models. Tab. 4 reports boldness accuracy for edits generated against each optimization model and evaluated across all three classifiers. The starred diagonal entries correspond to same-model evaluation, while the off-diagonal entries report evaluation on other classifiers. Three patterns emerge. First, same-model boldness accuracy is extremely low: ResNet-50 and ViT-B/16 both attain 0.0%, while CLIP ViT-L/14 attains only 12.7%. Second, the issue persists from the smaller ResNet-50 and ViT-B/16 to the larger CLIP ViT-L/14 foundation model. Third, boldness-attack samples transfer across models: edits optimized against one classifier also cause other classifiers to miss the verified target category, demonstrating cross-model transferability. **(4). Robustness-enhanced models still exhibit boldness failures.** The robustness-enhanced models exhibit two consistent findings. First, the evaluated robust checkpoints still exhibit failures. The ℓ_∞- and ℓ_2-robust ResNet-50 models attain only 4.5% and 4.4% boldness accuracy, respectively, while the AugMix model attains 3.7%; hence, models still fail on more than 95% of validated edits. This empirically confirms the distinction in Sec. 2: robustness to label-preserving changes does not imply correct adaptation after a label change. Second, this conclusion is consistent across different robustness-enhancement approaches. Both adversarially robust models and the AugMix model exhibit similarly low boldness accuracy, indicating that the observed gap is not specific to a single robustness method.

Table 3: FID on ImageNet (↓). "Rel. to Clean" is $\mathrm{FID}/\mathrm{FID}_{\text{clean–clean}}$. Per-image validity is reported in Tab. 2.

Method	FID ↓	Rel. to Clean
Clean–clean	5.6	1.00×
Nie et al. (2025)	12.7	2.27×
Ours	10.3	1.84×

Table 4: **Classification boldness accuracy (%).** Rows are optimization models; columns are evaluation models. * denotes same-model evaluation; clean accuracy is shown for reference.

Optimization model	Res-50	ViT-B	CLIP-L
Clean accuracy	88.7	89.1	91.2
Res-50	0.0*	47.2	57.3
ViT-B	42.6	0.0*	52.7
CLIP-L	46.3	47.1	12.7*

4.2 Object Detection

Experimental Settings for Detection **Datasets.** For object detection, we use 1,000 images sampled from the MS COCO 2017 validation set (Lin et al., 2014), which provides a multi-object setting with varied object scales, backgrounds, and spatial context, allowing us to evaluate localized category changes.

Evaluated Models. We evaluate four representative detector families: Faster R-CNN (Ren et al., 2015), a two-stage detector; SSD (Liu et al., 2016) and YOLOv3 (Redmon & Farhadi, 2018), one-stage detectors; and GLIP (Li et al., 2021), a language-grounded detector. We use publicly released pretrained COCO checkpoints. Faster R-CNN and SSD are loaded from TorchVision, while YOLOv3 and GLIP use their corresponding public implementations.

Detection Metrics. For each validated local edit, the reference annotation changes from (b_i, y_i) to (b_i, y_i'): the verified category changes to y_i', while b_i remains the reference spatial region. We evaluate semantic adaptation using *boldness mAP*, computed following the COCO protocol over IoU thresholds $\{0.50, 0.55, \ldots, 0.95\}$ against the edited annotations. A prediction therefore receives credit only if it both localizes the edited object and predicts its verified target category. To assess

whether edited objects remain spatially detected irrespective of category, we additionally report class-agnostic box recall at IoU 0.5. This follows the class-agnostic localization perspective in object detection (Jaiswal et al., 2020). Let X^s, $s \in \{\text{clean}, \text{edit}\}$, denote the clean or edited image and let B denote the reference box. We define $\text{BR}^s_{50} = \Pr[\exists\, \hat{b} \in \mathcal{P}(X^s) : \text{IoU}(\hat{b}, B) \geq 0.5]$, where $\mathcal{P}(X)$ denotes the predicted boxes with category labels ignored. Thus, BR_{50} measures whether the edited object continues to receive a spatially matched detection irrespective of its predicted category. We use one-to-one class-agnostic matching when multiple edited objects occur in the same image.

Local Edit Construction. For each eligible annotated object (b_i, y_i), we select a target category $y'_i \neq y_i$ and apply the local semantic-editing procedure described in Sec. 3.3. Pixels outside the edited region are preserved. We apply the per-instance validity protocol in Sec. C to the edited object region and retain only edits that satisfy the target-semantic and source-faithfulness criteria.



(a) Elephant (b) → Bear (c) Cat (d) → Dog

Figure 5: **Validated local semantic edits on COCO.** Edited objects acquire the target semantics while the detector retains the original categories.

Model	Clean mAP	Bold. mAP	$\text{BR}^{\text{clean}}_{0.5}$	$\text{BR}^{\text{edit}}_{0.5}$
Faster R-CNN	45.1	12.6	87.4%	76.6%
SSD	26.6	8.9	85.9%	75.3%
YOLOv3	34.3	10.6	92.1%	83.6%
GLIP	46.3	13.7	91.2%	83.7%

Table 5: **Object-detection boldness on validated COCO edits.** Clean and boldness mAP use original and verified edited annotations, respectively. $\text{BR}_{0.5}$ is class-agnostic box retention.

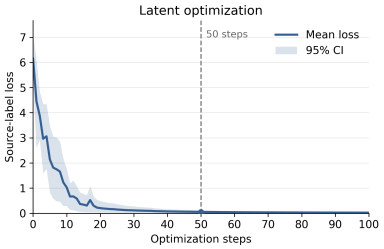
Experimental Results for Detection. Figure 5 illustrates local semantic edits on which the detector retains the original class. Tab. 5 compares clean and boldness mAP together with class-agnostic box retention. Two patterns emerge: **1. Boldness failures across detector families and model sizes.** All four evaluated detectors exhibit substantial mAP degradation under validated local edits. The failures span two-stage, one-stage, and language-grounded detectors, from smaller conventional models such as SSD to the larger multimodal GLIP model. Even GLIP, which has the highest clean mAP in this comparison, loses most of that performance after editing. Thus, the classification finding extends to localized predictions in multi-object scenes and is not confined to small models. **2. Category-adaptation failure despite retained localization.** After editing, class-agnostic box retention remains above 75% for every evaluated detector, with only a modest decrease from clean evaluation. Spatially matched detections therefore remain common despite the substantial loss in class-aware mAP. Together with the qualitative examples, this pattern supports category-adaptation failure as a contributor, rather than explaining the entire performance loss by the absence of localized detections.

4.3 Ablation Study

Table 6: **Source-regularization ablation.** Bold: default λ_z.

λ_z	DreamSim ↑	Valid (%)	BA (%)
0	0.51	88.7	0.0
10^{-3}	0.59	87.6	0.0
$\mathbf{3 \times 10^{-3}}$	**0.62**	**90.4**	**0.0**
10^{-2}	0.63	91.4	1.3
3×10^{-2}	0.61	90.1	3.2
10^{-1}	0.62	90.6	5.1

Figure 6: **Latent optimization.** Mean loss with 95% bootstrap CIs; dashed: 50 steps.



Source-Regularization Ablation. Tab. 6 shows the tradeoff between source preservation and attack strength. Around the default $\lambda_z = 3 \times 10^{-3}$, source similarity and validity remain high while boldness accuracy stays at 0%. Increasing λ_z further does not monotonically improve source similarity

or validity, but raises boldness accuracy, indicating that overly strong regularization weakens the attack. We therefore use $\lambda_z = 3 \times 10^{-3}$ as the default.

Optimization-Step Sensitivity. Figure 6 tracks source-label cross-entropy over 100 latent optimization updates on a fixed set of 100 images. Most of the reduction occurs during the early iterations, and the loss changes only marginally after 50 updates. This empirical stabilization supports 50 updates as a practical optimization budget for the examined samples.

5 Related Work

Robustness and Generative Attacks. Unrestricted attacks relax fixed pixel-space budgets through natural color transformations or generative priors (Yuan et al., 2022; Pan et al., 2025). DiffAttack uses source inversion, diffusion-latent optimization, and content preservation to induce misclassification (Chen et al., 2023). Such constructions share tools with our method, but our evaluation requires a verified, source-faithful oracle-label change rather than a label-preserving attack. Diffusion purification instead uses the generative model for defense (Nie et al., 2022). The closer prediction-persistence constructions are discussed in Sec. 2.1.

Generative Image Editing. GAN inversion and latent manipulation enable source-specific editing (Xia et al., 2022; Patashnik et al., 2021). Diffusion and flow-matching models support conditional image generation (Ho et al., 2020; Lipman et al., 2022; Dhariwal & Nichol, 2021; Ho & Salimans, 2022), with latent-diffusion and rectified-flow architectures supporting text-to-image synthesis (Rombach et al., 2022; Esser et al., 2024). Inversion and attention-control methods help preserve source content during semantic editing (Song et al., 2020; Mokady et al., 2023; Chen et al., 2025; Hertz et al., 2022), while FLUX.1 Kontext supports in-context generation and editing (Black Forest Labs et al., 2025).

Plug-in Optimization with Generative Priors. DMPlug, D-Flow, and FMPlug optimize the input of a fixed pretrained generator under task-specific objectives (Wang et al., 2024; Ben-Hamu et al., 2024; Wan et al., 2025). We adopt this principle for boldness, combining target conditioning with an original-label prediction loss and a source-relative latent penalty (Sec. 3).

6 Conclusions

We introduced *boldness*, the ability of a model to update its prediction after a valid nearby change alters the oracle label, complementing robustness under label-preserving changes. We formulated boldness evaluation through source-faithful semantic edits and developed a constrained plug-in optimization method using a fixed conditional generator. Our experiments reveal three main findings:

- **Boldness failures are widespread.** They occur across classifiers of different architectures and model scales, including a vision–language foundation model.
- **Robustness does not imply boldness.** Adversarially robust and AugMix-trained models still exhibit substantial failures under verified label-changing edits.
- **Boldness failure extends beyond classification.** Object detectors exhibit low boldness mAP despite high box retention, indicating insufficient category adaptation rather than localization loss.

These results motivate training and evaluation methods that explicitly promote adaptation to valid semantic changes, and extending boldness to other tasks and modalities.

AI Use Statement

Generative AI tools were used to assist with manuscript editing, organization, and readability. All AI-assisted content, analyses, and suggestions were reviewed and verified by the authors. The authors take responsibility for the final content of the paper.

References

Heli Ben-Hamu, Omri Puny, Itai Gat, Brian Karrer, Uriel Singer, and Yaron Lipman. D-flow: Differentiating through flows for controlled generation. In *International Conference on Machine*

Learning, 2024. URL https://api.semanticscholar.org/CorpusID:267770268.

Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, Sumith Kulal, Kyle Lacey, Yam Levi, Cheng Li, Dominik Lorenz, Jonas Müller, Dustin Podell, Robin Rombach, Harry Saini, Axel Sauer, and Luke Smith. Flux.1 kontext: Flow matching for in-context image generation and editing in latent space, 2025. URL https://arxiv.org/abs/2506.15742. Accessed 2026-02-28.

Ali Borji. A new kind of adversarial example. *ArXiv*, abs/2208.02430, 2022a.

Ali Borji. A new kind of adversarial example. *arXiv preprint arXiv:2208.02430*, 2022b.

Ali Borji. A new kind of adversarial example: Measuring the human–model gap, and its relationship to ood detection. *arXiv preprint arXiv:2607.22722*, 2026.

Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *International Conference on Computer Vision (ICCV)*, 2021.

Jianqi Chen, Hao Chen, Keyan Chen, Yilan Zhang, Zhengxia Zou, and Zhenwei Shi. Diffusion models for imperceptible and transferable adversarial attack. *arXiv preprint arXiv:2305.08192*, 2023. URL https://arxiv.org/abs/2305.08192.

Yinan Chen, Jiangning Zhang, Yali Bi, Xiaobin Hu, Teng Hu, Zhucun Xue, Ran Yi, Yong Li, and Ying Tai. Image inversion: A survey from gans to diffusion and beyond. *ArXiv*, abs/2502.11974, 2025. URL https://api.semanticscholar.org/CorpusID:276421784.

Francesco Croce, Maksym Andriushchenko, Vikash Sehwag, Edoardo Debenedetti, Nicolas Flammarion, Mung Chiang, Prateek Mittal, and Matthias Hein. Robustbench: a standardized adversarial robustness benchmark. *arXiv preprint arXiv:2010.09670*, 2020.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794, 2021.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020. URL https://api.semanticscholar.org/CorpusID:225039882.

Logan Engstrom, Andrew Ilyas, Hadi Salman, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL https://github.com/MadryLab/robustness.

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, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. Scaling rectified flow transformers for high-resolution image synthesis, 2024. URL https://arxiv.org/abs/2403.03206. Accessed 2026-02-28.

Stephanie Fu, Netanel Y. Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *ArXiv*, abs/2306.09344, 2023. URL https://api.semanticscholar.org/CorpusID:259171761.

Ian J. Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *CoRR*, abs/1412.6572, 2015.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Dan Hendrycks, Norman Mu, Ekin D. Cubuk, Barret Zoph, Justin Gilmer, and Balaji Lakshminarayanan. AugMix: A simple data processing method to improve robustness and uncertainty. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

Amir Hertz, Ron Mokady, Jay M. Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *ArXiv*, abs/2208.01626, 2022. URL https://api.semanticscholar.org/CorpusID:251252882.

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 *Neural Information Processing Systems*, 2017. URL https://api.semanticscholar.org/CorpusID:326772.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance. *arXiv preprint arXiv:2207.12598*, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and P. Abbeel. Denoising diffusion probabilistic models. *ArXiv*, abs/2006.11239, 2020. URL https://api.semanticscholar.org/CorpusID:219955663.

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. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 20349–20360, 2023. URL https://api.semanticscholar.org/CorpusID:257636562.

Jörn-Henrik Jacobsen, Jens Behrmann, Richard Zemel, and Matthias Bethge. Excessive invariance causes adversarial vulnerability. In *International Conference on Learning Representations*, 2019. URL https://arxiv.org/abs/1811.00401.

Ayush Jaiswal, Yue Wu, Pradeep Natarajan, and P. Natarajan. Class-agnostic object detection. *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 918–927, 2020. URL https://api.semanticscholar.org/CorpusID:227227809.

Cassidy Laidlaw, Sahil Singla, and Soheil Feizi. Perceptual adversarial robustness: Defense against unseen threat models. *ArXiv*, abs/2006.12655, 2021.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10955–10965, 2021. URL https://api.semanticscholar.org/CorpusID:244920947.

Tsung-Yi Lin, Michael Maire, Serge J. 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*, 2014. URL https://api.semanticscholar.org/CorpusID:14113767.

Yaron Lipman, Ricky T. Q. Chen, Heli Ben-Hamu, Maximilian Nickel, and Matt Le. Flow matching for generative modeling. *ArXiv*, abs/2210.02747, 2022. URL https://api.semanticscholar.org/CorpusID:252734897.

Wei Liu, Dragomir Anguelov, Dumitru Erhan, Christian Szegedy, Scott Reed, Cheng-Yang Fu, and Alexander C. Berg. SSD: Single shot MultiBox detector. In *European Conference on Computer Vision*, 2016. URL https://arxiv.org/abs/1512.02325.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *ArXiv*, abs/1706.06083, 2018.

Ron Mokady, Amir Hertz, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Null-text inversion for editing real images using guided diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6038–6047, 2023.

Weili Nie, Brandon Guo, Yujia Huang, Chaowei Xiao, Arash Vahdat, and Anima Anandkumar. Diffusion models for adversarial purification. In *International Conference on Machine Learning*, 2022. URL https://api.semanticscholar.org/CorpusID:248811081.

Xingyang Nie, Guojie Xiao, Su Pan, Biao Wang, Huilin Ge, and Tao Fang. A new type of adversarial examples. *arXiv preprint arXiv:2510.19347*, 2025.

Zhengzhao Pan, Hua Chen, and Xiaogang Zhang. Diffadvmap: Flexible diffusion-based framework for generating natural unrestricted adversarial examples (icml 2025 poster). https://icml.cc/virtual/2025/poster/43968, 2025. Accessed 2026-02-28.

Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-driven manipulation of StyleGAN imagery. In *IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 2085–2094, 2021. URL https://arxiv.org/abs/2103.17249.

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*, 2021. URL https://api.semanticscholar.org/CorpusID:231591445.

Joseph Redmon and Ali Farhadi. YOLOv3: An incremental improvement. *arXiv preprint arXiv:1804.02767*, 2018. URL https://arxiv.org/abs/1804.02767.

Shaoqing Ren, Kaiming He, Ross B. Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 39:1137–1149, 2015. URL https://api.semanticscholar.org/CorpusID:10328909.

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.

Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *ArXiv*, abs/2010.02502, 2020. URL https://api.semanticscholar.org/CorpusID:222140788.

Stability AI. stabilityai/stable-diffusion-xl-base-1.0 model card. https://huggingface.co/stabilityai/stable-diffusion-xl-base-1.0, 2023. Accessed 2026-02-28.

Yonglong Tian, Dilip Krishnan, and Phillip Isola. Contrastive multiview coding. In *European Conference on Computer Vision*, pp. 776–794. Springer, 2020.

Florian Tramèr, Jens Behrmann, Nicholas Carlini, Nicolas Papernot, and Jörn-Henrik Jacobsen. Fundamental tradeoffs between invariance and sensitivity to adversarial perturbations. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 9561–9571, 2020. URL https://proceedings.mlr.press/v119/tramer20a.html.

Yuxiang Wan, Ryan Devera, Wenjie Zhang, and Ju Sun. Saving foundation flow-matching priors for inverse problems, 2025. URL https://openreview.net/forum?id=qqwx1hGwSE.

Hengkang Wang, Xu Zhang, Taihui Li, Yuxiang Wan, Tiancong Chen, and Ju Sun. DMPlug: A plug-in method for solving inverse problems with diffusion models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=81IFFsfQUj.

Weihao Xia, Yulun Zhang, Yujiu Yang, Jing-Hao Xue, Bolei Zhou, and Ming-Hsuan Yang. GAN inversion: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2022. doi: 10.1109/TPAMI.2022.3184310. URL https://arxiv.org/abs/2101.05278.

Shengming Yuan, Qilong Zhang, Lianli Gao, Yaya Cheng, and Jingkuan Song. Natural color fool: Towards boosting black-box unrestricted attacks. *ArXiv*, abs/2210.02041, 2022. URL https://api.semanticscholar.org/CorpusID:252715759.

Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2018.

A Pixel-Space Stress Tests and Semantic Validity

Pixel-space baseline. We use an implementation of the large-distance construction of Nie et al. (2025) as the pixel-space baseline. Its objective in Eq. (2) minimizes the original-label loss subject to a lower bound on the distance from the original image.

Why semantic validation remains necessary. The distance constraint specifies how far a candidate must move, but does not enforce image validity, label change, or source faithfulness. A low original-label loss therefore does not establish a valid boldness failure. Candidates from this baseline require the same semantic checks as generated edits before they can be used for validity-conditioned boldness evaluation.

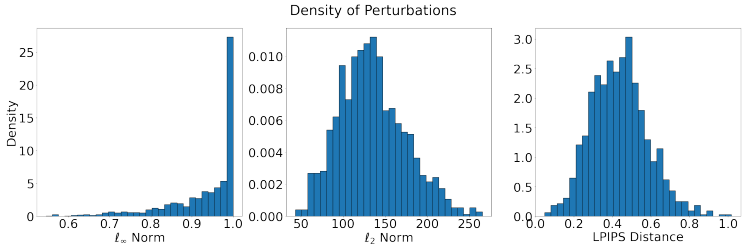


Figure 7: Distributions of ℓ_∞, ℓ_2, and LPIPS distances between original ImageNet images and their validated semantic boldness edits. The broad spread of pixel-space distances indicates that valid semantic transitions cannot be reliably characterized by a single fixed ℓ_p budget.

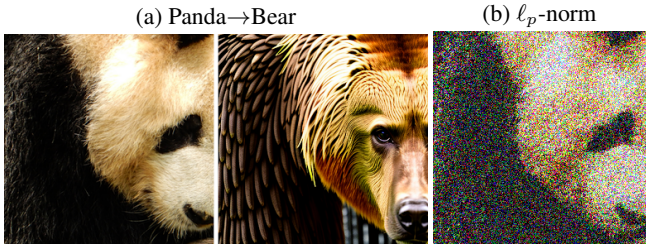


Figure 8: Comparison between semantic boldness editing and a large-budget ℓ_p-based stress test. The left panel shows a source image and its source-faithful semantic edit from panda to bear. The right panel shows the large-budget pixel-space result, which is dominated by visual artifacts and does not establish a valid semantic transition.

Distances of Validated Semantic Edits Figure 7 reports the distributions of ℓ_∞ (left), ℓ_2 (middle), and LPIPS (right) distances between original ImageNet images and their validated semantic edits. The ℓ_∞ distances range approximately from 0.5 to 1.0, while the ℓ_2 distances span roughly 50 to 250. These distributions describe the magnitude of edits that pass the validity protocol. They do not themselves establish a label change: a pixel-space threshold can restrict the accepted displacement scale, but cannot replace checks of target semantics, source faithfulness, and image validity.

B Gaussian-Shell-Constrained Latent Optimization

Constraint and scope. The source regularizer discourages departure from the inversion seed, whereas the shell limits radial departure from the Gaussian prior's concentration region. The two restrictions therefore serve different roles. For completeness, the shell used in Sec. 3 is

$$\mathcal{S}_\epsilon := \left\{ z \in \mathbb{R}^m : (1-\epsilon)\sqrt{m} \le \|z\|_2 \le (1+\epsilon)\sqrt{m} \right\}, \tag{8}$$

where $0 < \epsilon < 1$ and m is the seed dimension. This constraint is used here because the optimized variable is the sampler's initial Gaussian noise latent. It should not be applied unchanged to prompt embeddings or clean VAE image latents. The shell controls the seed norm, not the complete seed distribution or per-image validity.

Projected updates. We solve Eq. (7) by projected gradient descent, initialized at the projected source inversion seed:

$$z^{(0)} = \Pi_{\mathcal{S}_\epsilon}(z^*). \tag{9}$$

We then alternate a gradient step and a projection onto the shell:

$$z^{(k+1/2)} = z^{(k)} - \eta \nabla_z \left[\ell\Big(f\Big(G_\theta(z^{(k)}; c_{y'}) \Big), y \Big) + \lambda_z \frac{\|z^{(k)} - z^*\|_2^2}{m} \right], \tag{10}$$

$$z^{(k+1)} = \Pi_{\mathcal{S}_\epsilon}\Big(z^{(k+1/2)} \Big). \tag{11}$$

For $z \neq 0$, the Euclidean projection onto $\mathcal{S}_\epsilon$ has the closed form

$$\Pi_{\mathcal{S}_\epsilon}(z) = \begin{cases} (1+\epsilon)\sqrt{m}\, z/\|z\|_2, & \|z\|_2 > (1+\epsilon)\sqrt{m}, \\ (1-\epsilon)\sqrt{m}\, z/\|z\|_2, & \|z\|_2 < (1-\epsilon)\sqrt{m}, \\ z, & \text{otherwise.} \end{cases} \tag{12}$$

For $z = 0$, the Euclidean projection is non-unique; any vector with norm $(1-\epsilon)\sqrt{m}$ is a valid projection.

C Experiment details

The optimization searches for prediction failures, whereas the validity protocol checks whether a candidate is an appropriate semantic edit. We assess the following three properties before evaluating boldness.

1. **Target semantic correctness.** We use TIFA (Hu et al., 2023) to assess consistency with the intended target semantics y', denoting its per-image score by S_{target}.
   Given the target description, TIFA generates a set of structured questions whose answers must be consistent with the generated image. A pretrained VQA model is then used to answer these questions, and we define the semantic faithfulness score as
   $$S_{\text{target}} = \frac{\#\text{correct VQA answers}}{\#\text{questions}}.$$
   A high value of S_{target} indicates that the perturbed image faithfully reflects the intended target concept.
2. **Source faithfulness.** We assess whether the edited image remains visually related to the original image rather than relying on target semantics alone. To quantify this property, we measure perceptual similarity between the original and perturbed images using DreamSim embeddings $h_{\text{DS}}(\cdot)$:
   $$S_{\text{src}}(x, x') = \text{sim}(h_{\text{DS}}(x), h_{\text{DS}}(x'))\,,$$
   where $\text{sim}(\cdot,\cdot)$ denotes cosine similarity in a perceptual embedding space. We use DreamSim (Fu et al., 2023) as the source-faithfulness metric in all reported experiments. LPIPS (Zhang et al., 2018) and DINO similarity (Caron et al., 2021) provide alternative perceptual measures, but are not used for the validity threshold reported here.
3. **Editing vs. random generation.** Target conditioning can produce a plausible target-class image without retaining the particular input. We therefore compare each edit's source similarity with a random-generation baseline using the same target prompt. For an original image x, let x'_{edit} be the edit and $\{x'_{\text{rand},k}\}_{k=1}^K$ the K random target generations. We compute
   $$\text{Gap} = S_{\text{src}}(x, x'_{\text{edit}}) - \frac{1}{K}\sum_{k=1}^{K} S_{\text{src}}(x, x'_{\text{rand},k}).$$
   A positive gap means that the edit is more source-similar than the average of these random target generations.

Acceptance rule. An edit is retained only if all three scores satisfy the thresholds in Tab. 2. The overall validity pass rate is 90.4%. Only retained edits enter the boldness evaluation; passing the validity filter is not itself a prediction failure.