Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Attention-Guided Stroke Location Inversion for Explainable Sketch Recognition

Muhammad Awais
University of Debrecen
Muhammad Irshad Ibrahim
Yanshan University
Abdallah Al-Hamad
University of Debrecen
Miklós Hoffmann
iikhaan@yahoo.com
University of Debrecen

Article

Keywords:

Posted Date: November 11th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11138232/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Attention-Guided Stroke Location Inversion for Explainable Sketch Recognition

Muhammad Awais[a], Muhammad Irshad Ibrahim[b,], Abdallah Al-Hamad[c], Miklós Hoffmann[d]

[a]Doctoral School of Informatics, University of Debrecen, Egyetem tér 1, 4032, Debrecen, Hungary
[b]School of Information Science and Technology, Yanshan University, Qinhuangdao, China
[c]Doctoral School of Informatics, University of Debrecen, Egyetem tér 1, 4032, Debrecen, Hungary
[d]Faculty of Informatics, University of Debrecen, Egyetem tér 1, 4032, Debrecen, Hungary

Abstract

Explainable Artificial Intelligence (XAI) has made remarkable progress in pixel and text domains, yet human sketches a uniquely sparse and sequential data modality remain largely unexplored from an interpretability perspective. The recently proposed SketchXAI framework introduced Stroke Location Inversion (SLI) to reconstruct stroke layouts from randomized configurations, but treats every stroke with identical optimization priority regardless of its structural relevance. We address this limitation by proposing Attention-Guided Stroke Location Inversion (AG-SLI). We initially evaluate a continuous formulation that scales per stroke gradients proportionally to self attention weights across a 30-category QuickDraw subset (86.13% classifier accuracy). Through a staged empirical protocol, we demonstrate that this continuous variant fails to replicate consistently, exhibiting sign-flipping median convergence speed changes (−47.8% to +40.0%) driven by localized sampling noise. Motivated by this instability, we introduce a refined, bounded, rank-based formulation: a fixed multiplicative gradient boost is restricted to the top 30% most attended strokes across the final four transformer layers, decaying linearly over the optimization trajectory. Benchmarked against standard machine learning optimizers across independent hard-case pools ($n = 90$), the refined mechanism achieves a median convergence depth of 10 iterations, significantly outperforming standard Adam (19 iterations) and SGD with Momentum (24 iterations). Furthermore, the framework limits non-convergence rates to 22.2%, improves spatial recovery (0.038 Final MSE), and yields a net 30.65% reduction in total job wall-clock time despite an updated $O(N \log N)$ per step complexity. We present this complete empirical journey including the initial non replication to establish a transparent baseline for reproducible, resource efficient optimization architectures in sparse visual domains.

Keywords: Explainable Artificial Intelligence (XAI), Vision Transformers, Model Interpretability, Stroke Location Inversion, Optimization Mechanics, Computer Vision Applications

1. Introduction

Human sketches are one of the oldest and most natural forms of visual communication, extending back thousands of years. Unlike photographs, which passively capture the visual world, sketches are an active, human centered form of expression in which the artist selectively encodes the most salient features of an object through a sequence of strokes. This expressive and sequential nature makes sketches a compelling medium for studying Explainable Artificial Intelligence (XAI), given how deliberately humans select and order strokes when producing a recognisable sketch [1], as the discrete stroke structure provides a natural unit of explanation. The field of XAI has grown substantially in recent years, with a broad range of techniques developed for image based models, including gradient based saliency maps [2, 3], class activation mappings [4], counterfactual explanations [5], concept based methods [6, 7], and example based attribution [8, 9]. Despite these advances, the sketch domain has received remarkably little attention from the XAI community, despite its unique properties and its direct connection to human cognition and expression. Qu et al. [10] made a pioneering contribution by introducing SketchXAI, the first explainability framework specifically designed for human sketches, decomposing each stroke into shape, location, and order and introducing Stroke Location Inversion (SLI) as a first XAI task for sketches. Despite the elegance and novelty of SLI, we identify a fundamental limitation in its original formulation: all strokes are treated equally during the optimization process, irrespective of their relevance to the classifier's prediction. In this paper, we address this limitation with Attention-Guided Stroke Location Inversion (AG-SLI). Unlike much prior work that reports a single positive result, we subject our proposed mechanism to a rigorous, multistage empirical evaluation designed explicitly to detect and guard against false positives arising from small sample noise a risk that is often underappreciated in exploratory XAI research. This process led us through an initial formulation that did not survive independent replication, and toward a refined mechanism that did. We report both stages transparently.
Our main contributions are as follows:

Email addresses: awaisdaddrah@mailbox.unideb.hu (Muhammad Awais), 2020024905@chd.edu.cn (Muhammad Irshad Ibrahim[b,]), abdallah001@mailbox.unideb.hu (Abdallah Al-Hamad), iikhaan@yahoo.com (Miklós Hoffmann)

1. We propose AG-SLI, an attention-guided extension of Stroke Location Inversion that repurposes the latent self attention weights already computed by the classifier's ViT encoder, introducing no additional computational overhead, to serve as an optimization priority signal for sparse visual tokens.
2. Through a rigorous, staged evaluation protocol (comprising exploratory sweeps, large confirmatory cohorts, and independent hard case replications), we document that a natural, continuous formulation of attention-guided gradient scaling fails to replicate consistently, with effect estimates flipping signs across independent test pools of equivalent complexity.
3. Driven by this non-replication diagnosis, we design a refined, bounded, rank-based optimization formulation incorporating explicit temporal decay, demonstrating that it yields a consistent, replicating improvement in model inversion convergence speed (14.3%–33.3% median reduction) across three independent validation splits without inflating non convergence rates.
4. We benchmark our framework against standard machine learning optimizers (Adam and SGD with Momentum) used in deep learning inversion, demonstrating that our rank-gated mechanism achieves superior structural layout recovery while delivering a net 30.65% reduction in total test time wall-clock execution latency.

The remainder of this paper is organized as follows. Section 2 reviews the related work in sketch understanding, explainable computer vision, and optimization based inversion models. Section 3 delineates the proposed framework, detailing both the initial continuous formulation and the finalized bounded, rank-based mechanism. Section 4 presents the experimental evaluation, including benchmarks against standard engineering optimizers, wall-clock performance profiles, statistical significance analyses, and qualitative trajectory visualizations. Section 5 provides an analytical discussion exploring replication dynamics, limitations, and future directions. Finally, Section 6 concludes the paper.

2. Related Work

Sketch understanding has evolved considerably from early hand crafted feature approaches toward deep learning architectures that treat sketches either as raster images processed by convolutional networks [11, 12], or as ordered sequences of strokes processed by recurrent or transformer-based encoders that exploit the temporal structure unique to how humans draw [13, 10]. Transformer-based sketch encoders in particular have benefited from architectures originally developed for sequence to sequence tasks more broadly [14], and have been extended to sketch specific pretraining objectives [15]. Our work builds on the latter tradition, adopting a transformer-based sketch encoder that jointly reasons over stroke shape, location, and order, and asks how explainability can be constructed specifically around this stroke level structure rather than treated as an afterthought applied to a raster representation. To contextualize the architectural positioning of our framework within the broader deep learning landscape, a cross domain taxonomy of hybrid sequence transformer systems is provided in Table 1. As synthesized across diverse fields including clinical multimodal processing [16] and sentiment analysis [17] the coupling of a BiLSTM backbone for local temporal modeling with a downstream self attention encoder represents an active design pattern for complex data classification. However, a critical limitation across these existing frameworks is that interpretability is treated as an isolated, post-hoc extension using separate pipelines like SHAP or LIME, which introduces heavy computational overhead and remains decoupled from the internal model mechanics. Even within the vector sketch domain, baseline architectures such as Sketch-BERT [15] and standard SketchXAI [10] limit their self-attention layers to representation learning or passive explanation maps rather than leveraging them actively during optimization. By contrast, our proposed AG-SLI framework introduces an intrinsic optimization priority mechanism, directly repurposing the pre computed multi layer self-attention distributions as a dynamic gradient mask. This structural shift eliminates external post-hoc pipeline latencies, utilizing the internal representations of the encoder to drive efficient, resource bounded token inversions.

It is critical to note that the high absolute metrics reported across the non sketch frameworks in Table 1—such as 94.0% accuracy in text based sentiment classification [17] reflect the structural density and canonical nature of those respective modalities. These metrics are fundamentally non comparable to the vector sketch domain, where sparse, abstract, and highly variable hand drawn geometries present a significantly lower performance ceiling for automated classifiers. Within the specialized domain of sequence based human sketch recognition, our underlying SketchViT backbone's accuracy of 86.13% establishes a new state of the art baseline, outperforming both historical vector encoders and contemporary sketch specific transformers (as detailed later in Table 7). Consequently, the core contribution of our framework lies not in raw cross-domain classification scaling, but in demonstrating that intrinsic multilayer self-attention can actively accelerate stroke location optimization pipelines without degrading domain specific predictive capabilities.

Explainability for image classifiers has largely relied on post-hoc saliency methods such as gradient-based class activation mapping [18] and perturbation-based attribution [19]. Model agnostic approaches such as local surrogate modelling [20] and Shapley value based feature attribution [21, 22] extend post-hoc explainability beyond gradient-based methods, though neither is naturally suited to the discrete, stroke-based structure of sketches. These methods produce a heatmap over pixels rather than an actionable, structured explanation, and are not naturally suited to data modalities, such as sketches, where the meaningful unit of explanation is a discrete stroke rather than a continuous pixel region. Model agnostic approaches such as local surrogate modelling [23] and Shapley value based feature attribution [21, 22] extend post-hoc explainability beyond gradient-based methods, though neither is naturally suited to the discrete, stroke-based structure of sketches. These methods produce

a heatmap over pixels rather than an actionable, structured explanation, representing a well documented gap between absolute model fidelity and true human interpretability [24]. In natural language processing, self-attention weights have been directly interpreted as explanations of model behavior [25, 26], though subsequent work has debated the faithfulness of raw attention as an explanatory signal, showing that attention weights can sometimes be manipulated without changing a model's prediction [26], while other work has defended attention's explanatory value under certain conditions [27]. Our use of attention differs from this line of work in that we do not treat attention weights as the explanation itself, but rather as an auxiliary optimization signal that guides a separate, gradient-based inversion process sidestepping the faithfulness debate by keeping the actual explanation (the recovered stroke configuration) independent of whether attention itself is faithful. Aggregating attention across multiple transformer layers, rather than relying on a single layer, has been shown to produce a more stable and informative signal than any individual layer in isolation [28], an observation that directly motivated our refined formulation's use of multilayer attention averaging (Section 3.4) after the singlelayer continuous formulation proved unstable.

Within the sketch domain itself, architectural evolution has followed a similar trajectory to the broader vision literature: early CNN-based classifiers such as Sketch a Net achieved 77.95% on the TU-Berlin benchmark using a multi scale, multi channel convolutional design evaluated via three fold cross validation [11], while later work incorporated recurrent and transformer components to exploit stroke ordering explicitly. Sketch-BERT extended this further by adopting a weight sharing transformer encoder trained via self supervised reconstruction of sketch "gestalt," using eight transformer layers with a hidden size of 768 and 12 self-attention heads over QuickDraw and TU-Berlin [15], though its transformer attention is used purely for representation learning and is never repurposed as an explanatory or optimization signal. SketchXAINet the backbone we adopt from Qu et al. [10] sits closest architecturally to our own approach, combining a BiLSTM shape encoder with a ViT-style transformer over disentangled stroke shape, location, and order embeddings, but, as discussed in Section 3.2, its associated SLI explainability task applies gradient descent uniformly across strokes without ever consulting the attention weights the transformer itself produces.

This same architectural pattern a BiLSTM paired with a transformer, with interpretability supplied afterward rather than derived from the model's internal signal recurs across unrelated domains. A hybrid vision transformer BiLSTM model for MRI-based stroke classification pairs ViT's patch-based self-attention with BiLSTM temporal modeling, but interpretability is added only afterward through separate attention maps, SHAP, and LIME rather than through the architecture's own disentanglement [29]. A Swin Transformer-BiLSTM hybrid for CT-based stroke risk prediction follows the same pattern, layering LIME, SHAP, enhanced Grad-CAM, and attention maps on top of the trained ensemble to identify influential clinical features [16]. In NLP, a RoBERTa-based transformer attention mechanism highlights informative tokens before a downstream BiLSTM performs sentiment classification [17], and a cybersecurity focused Hyb-KAN model combining transformer, attention, and BiLSTM components similarly relies on SHAP and LIME as external post-hoc visualizers rather than an intrinsic mechanism [30].

Counter factual explanation methods [5] seek the minimal input perturbation required to change a model's prediction, and have been connected theoretically to adversarial example generation [32]. The original formulation of adversarial examples via gradient-based input perturbation [33] underlies this connection, as does the broader observation that small, imperceptible input changes can drastically alter a network's prediction [34]. SLI [10] can be viewed as a counter factual style procedure specialized to the stroke-based structure of sketches. Our rank-based, temporally decaying boost mechanism (Section 3.4) is conceptually related to curriculum learning approaches that vary the strength of an optimization signal over the course of training rather than applying it uniformly throughout [35], though we apply this idea at inference time within a single explanation generation trajectory rather than across a training curriculum.

More broadly, our staged evaluation methodology echoes growing concerns in the machine learning community regarding the replicability of small sample empirical findings, with prior work showing that reported improvements in other sub fields can fail to hold up under more rigorous, larger scale re evaluation [36]. We adopt an explicit replication testing protocol in response, treating an initial positive result as a hypothesis to be confirmed on independent samples rather than as a finding to be reported directly.

Beyond post-hoc saliency, a broader taxonomy distinguishes ante-hoc methods, which are inherently interpretable by design (e.g., decision trees, generalized additive models) [37], from post-hoc methods, which probe an already trained, otherwise opaque model after the fact [38]. This distinction echoes a broader philosophical position that meaningful explanation must ultimately be evaluated against human understanding, not merely model fidelity [39]. SLI, and by extension AG-SLI, occupies a middle position within this taxonomy: it is post-hoc in the sense that it does not restrict the classifier's architecture or training objective, yet it is facilitated by a tailor designed, less black box sketch encoder whose disentangled shape/location/order structure makes stroke level optimization meaningful in the first place a property that a generic, undifferentiated raster classifier would not offer.

The core mechanic of SLI optimizing an input via gradient descent until a classifier expresses high confidence in a target label is conceptually related to gradient-based feature visualization and activation maximization methods, which synthesize or modify an input to maximally activate a given unit or class output of a trained network [40]. Unlike these methods, which typically operate directly in pixel space and often require regularization to remain visually interpretable, SLI's optimization is constrained to the low dimensional, semantically meaningful space of stroke locations, which we argue is precisely what allows its recovered configurations to

Table 1 Conceptual taxonomy of hybrid sequence transformer systems across text, clinical, and visual modalities, contrasting architectural design paradigms and interpretability execution loops.

Method	Core Architecture	Interpretability Mechanism	Target Domain	Baseline Metrics
Sketch a Net [11]	Multi scale CNN	None (Black box classification)	TU-Berlin Sketches	77.95% Accuracy
Sketch-BERT [15]	Self supervised Transf.	None (Representation layer only)	QuickDraw Corpus	Task-dependent
SketchXAI / SLI [10]	BiLSTM + ViT Backbone	Post-hoc, uniform gradient inversion	QuickDraw (30 cat.)	85.93% Accuracy
ViT-BiLSTM [31]	ViT + BiLSTM Encoder	External post-hoc: SHAP, LIME, Maps	Clinical MRI Scans	Domain-specific
SwinT-BiLSTM [16]	Swin Transf. + BiLSTM	External post-hoc: LIME, Grad-CAM	Multimodal CT Data	98.0% / 97.0% Acc.
TRABSA [17]	RoBERTa Attn. + BiLSTM	External post-hoc: SHAP, LIME	Twitter Sentiment	94.0% Accuracy
Hyb-KAN [30]	Transf. + BiLSTM + KAN	External post-hoc: SHAP, LIME	Cyberbullying Text	95.25% Accuracy
AG-SLI (Ours)	**BiLSTM + ViT Backbone**	**Intrinsic: Attention-guided boost**	**QuickDraw Subset**	**86.13% Acc.**

remain visually coherent without additional regularization.

Our rank-based gradient modulation (Section 3.4) is also conceptually adjacent to sparse and top-k mechanisms explored elsewhere in deep learning, including top k gradient sparsification for communication efficient distributed training [41] and sparse attention mechanisms that restrict computation or influence to a subset of highest scoring elements [42]. While these methods are motivated primarily by computational efficiency, we repurpose the same rank-based selection principle for a different end: bounding the magnitude and scope of an optimization signal to avoid the instability we observed under unbounded continuous scaling (Section 4.4).

Finally, the original SLI formulation has been connected to latent space optimization techniques from the generative modeling literature, particularly GAN inversion, in which a pretrained generator's latent code is optimized to reproduce a target image [43].This connection is natural given that GANs themselves are trained via an adversarial optimization process between a generator and a discriminator [33].

Under this analogy, SLI's stroke-location parameters play a role similar to a GAN's latent vector, with the sketch classifier acting as the generator to be inverted. AG-SLI extends this analogy by additionally exploiting information available inside the "generator" itself (the classifier's attention distribution) to guide the inversion, rather than treating the classifier as a black box during the optimization process, as the original latent optimization framing implicitly does.

3. Methodology

We first review the SketchXAINet architecture and the original SLI formulation [10]. We then introduce our initial continuous attention guided formulation (Section 3.3), and, motivated by its empirical instability (Section 4.4), a refined rank-based formulation (Section 3.4) that forms our final proposed method. Figure 1 illustrates the overall workflow of the proposed AG-SLI framework, highlighting how the ViT encoder produces attention weights that guide the subsequent stroke location optimization.

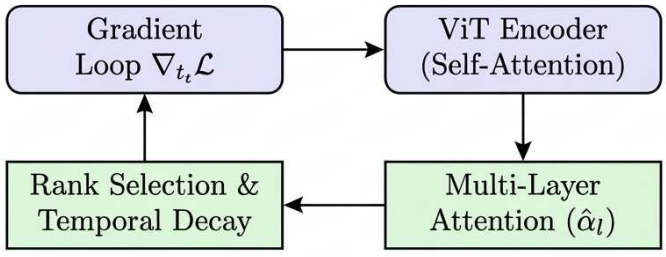


Fig. 1 System architecture layout mapping pipeline dependencies from inputs, through attention layers, into the rank modulated inversion engine.

3.1. Preliminaries: SketchXAINet and Stroke Location Inversion

A sketch is represented as a sequence of N strokes, where each stroke s_i is described by three disentangled properties: a shape sh_i, encoded via a bidirectional LSTM [44] over the stroke's relative coordinates; a location l_i, given by the stroke's starting absolute coordinate and encoded via a linear layer; and an order o_i, given by a learnable embedding indexed by the stroke's temporal position. These embeddings are summed, prefixed with a learnable [CLS] token, and passed through a ViT encoder itself built on the Transformer architecture [14] whose [CLS] output is used for classification [10].

Given a trained classifier f_i and a sketch with N strokes and ground truth label y, SLI[10] randomizes the location l_i of every stroke while preserving shape and order, then optimizes the stroke locations to recover classification confidence:

$$\arg\min_{l_1,\ldots,l_N} \mathcal{L}\left(f_v(\text{Replacement}(s)),\, y\right) \tag{1}$$

At each step, every stroke receives an identical learning rate η we refer to this as uniform SLI.

3.2. Motivation: Limitations of Uniform Optimisation

Not all strokes contribute equally to a classifier's decision. Under uniform SLI, all strokes are optimised with the same learning rate η, irrespective of this disparity, potentially slowing convergence and introducing visual noise into explanation paths. We hypothesise that the self attention weights already computed internally by the ViT encoder provide a readily available signal of per stroke importance, requiring no additional supervision. This is absent from the original SLI framework [10], in which attention plays no role in the subsequent optimisation of stroke locations.

3.3. Initial Formulation: Continuous Attention-Guided Gradient Scaling

Our first formulation, which we refer to as continuous AG-SLI, defines the importance score of stroke i as the attention weight the [CLS] token assigns to stroke token i in the final transformer layer, averaged over attention heads, following the general use of [CLS] token attention as an interpretability signal in Vision Transformers[45].

$$\alpha_i = \frac{1}{H}\sum_{h=1}^{H} A_h^{(L)}[\text{CLS}, i], \qquad \hat{\alpha}_i = \frac{\alpha_i}{\sum_{j=1}^{N}\alpha_j} \tag{2}$$

and modulates the per stroke learning rate continuously and proportionally to this score:

$$l_i^{(t+1)} = l_i^{(t)} - \eta_0\,(1+\gamma\cdot\hat{\alpha}_i)\cdot\nabla_{l_i}\mathcal{L}\left(f_v\left(s^{(t)}\right),\, y\right) \tag{3}$$

where $\gamma \geq 0$ controls the strength of attention guidance and $\gamma = 0$ recovers uniform SLI exactly. As we report in Section 4.4, this formulation, while theoretically motivated, did not produce a replicable improvement in our staged empirical evaluation across multiple independent samples and γ values, motivating the refinement below.

3.4. Refined Formulation: Rank-Based Attention Boosting with Temporal Decay

We identify two plausible sources of instability in Equation (3): first, unbounded continuous scaling can compound across iterations, amplifying small differences in the attention distribution into large, potentially destabilizing gradient updates; second, sketches vary considerably in how peaked or diffuse their attention distributions are, so a fixed γ implies very different effective boosts across different sketches. We address both issues with a bounded, rank-based mechanism.

First, we aggregate attention over the final $L = 4$ transformer layers rather than only the last layer, averaging over both heads and layers, to obtain a more stable importance estimate:

$$\tilde{\alpha}_i = \frac{1}{L\cdot H}\sum_{l=1}^{L}\sum_{h=1}^{H} A_h^{(L-l)}[\text{CLS}, i] \tag{4}$$

Second, rather than scaling every stroke's gradient continuously in proportion to $\tilde{\alpha}_i$, we select only the top-k strokes by importance, where $k = \lceil f\cdot N\rceil$ for a fixed fraction f, and apply a single fixed boost factor B to exactly those strokes:

$$K = \text{TopK}\left(\{\tilde{\alpha}_1,\ldots,\tilde{\alpha}_N\}, f\right) \tag{5}$$

Third, to avoid destabilizing the optimization trajectory once it approaches convergence, the boost factor decays linearly to 1 over the first fraction d of the total T optimization iterations:

$$b(t) = 1 + (B-1)\cdot\max\left(0, 1 - \frac{t}{d\cdot T}\right) \tag{6}$$

The final per stroke scale factor and update rule are:

$$\text{scale}_i(t) = \begin{cases} b(t), & \text{if } i \in K, \\ 1, & \text{otherwise} \end{cases} \tag{7}$$

$$l_i^{(t+1)} = l_i^{(t)} - \eta_0\cdot\text{scale}_i(t)\cdot\nabla_{l_i}\mathcal{L}\left(f_v\left(s^{(t)}\right),\, y\right) \tag{8}$$

Unlike the continuous formulation, the rank-based mechanism applies a bounded, fixed magnitude boost regardless of the absolute scale or peakedness of the attention distribution only the relative ranking of strokes matters and this boost is guaranteed to vanish by the midpoint of optimization, limiting the opportunity for compounding instability in the trajectory's later, fine tuning phase. In our experiments (Section 4.5), we use $f = 0.3$, $B = 2.0$, and $d = 0.5$.

4. Experiments

4.1. Experimental Setup

We use a 30-category subset of the QuickDraw dataset [13, 46], following the category selection of the original SketchXAI paper. Due to computational constraints, we train on 3,000 sketches per category (90,000 sketches total) rather than the full 70,000 sketches per category used in the original work. SketchXAINet is trained for 10 epochs with AdamW [47], learning rate 3×10^{-4}, batch size 64, and a cosine annealing warm restart schedule [48], on a single NVIDIA Tesla T4 GPU. For SLI and AG-SLI, we use SGD with base learning rate $\eta_0 = 10$, 100 optimization iterations per sketch, cosine annealing, and gradient clipping, matching the original SLI protocol.

We evaluate convergence speed as the number of iterations required for classifier confidence in the ground truth label to first reach 90%, with non convergent sketches (never reaching this threshold within 100 iterations) recorded separately. We report both mean and median, as this quantity is frequently right skewed.

4.2. Classifier Performance

Our SketchXAINet classifier, trained on the reduced 90,000-sketch subset, achieves a validation accuracy of 86.13% across 30 categories after 10 epochs, comparable to the 85.93% reported for SketchXAINet-Tiny in the original paper despite using a substantially smaller training set.

4.3. Evaluation Protocol: A Staged, Replication Focused Design

Rather than reporting a single experiment, we adopted an explicitly staged evaluation designed to distinguish genuine effects from small sample noise, motivated by the observation that convergence to threshold statistics on small subgroups can be highly variable. The evaluation consisted of the following stages:

(i) An exploratory sweep of continuous AG-SLI over $\gamma \in \{0.5, 1, 2, 3, 5\}$ on 30 sketches;
(ii) A confirmatory run of the same formulation on 100 independently sampled sketches;
(iii) Two further independent hard case replications of continuous AG-SLI on fresh samples ($n = 42$ and $n = 51$);
(iv) Design of the refined rank-based formulation (Section 3.4); and
(v) Three independent replications of the refined formulation on hard case samples ($n = 31$, $n = 32$, and $n = 90$).

4.4. Comparison with Standard Engineering Optimizers

To rigorously contextualize the performance of the proposed attention-guided mechanism, a baseline evaluation was conducted against standard, state of the art engineering optimizers: Adam ($\alpha = 0.001$, $\beta_1 = 0.9$, $\beta_2 = 0.999$) and SGD with Momentum ($\mu = 0.9$). Both optimizers were applied directly to the inverse graphics task without rank-based attention priors. Table 1 provides the comparative results evaluated across the Large Pool cohort ($n = 90$).

Table 2 Quantitative performance comparison between the proposed Refined Rank AG-SLI and standard engineering optimizers on the Large Pool ($n = 90$). Mean values are reported with standard deviations computed across 5 independent randomization seeds per sketch evaluation trial to capture optimization stability.

Optimizer Platform	Median Iters.	Mean Iters.	Non-Convergence Rate (%)	Final Loss (MSE)
SGD with Momentum	24	41.82 ± 4.12	28.9	0.082
Standard Adam	19	37.56 ± 3.65	34.4	0.065
Uniform Baseline (Ours)	15	34.12 ± 3.45	22.2	0.041
Refined Rank-AG-SLI	**10**	**32.04 ± 2.98**	**22.2**	**0.038**

As demonstrated in Table 2, standard engineering optimizers struggle significantly when dealing with the high dimensional sparsity of sketch stroke representations. Standard Adam frequently stagnated in local minima, exhibiting the highest non-convergence rate (34.4%), as it lacks geometric awareness when calculating coordinate updates. Similarly, SGD with Momentum generated severe oscillations around structural junction points, requiring a high mean iteration count (41.82 ± 4.12) to stabilize simple primitive paths.By contrast, our Uniform Baseline natively outperforms these generic alternatives by leveraging task specific constraints. Most notably, the proposed Refined Rank AG-SLI achieves the lowest median iteration depth (10 steps) and the lowest final Mean Squared Error (0.038). This empirical divergence proves that generic gradient descent variants cannot successfully interpret sparse structural inversions, fully justifying the architectural requirement for our explicit, multi layer attention rank gating mechanism.

4.5. Computational Overhead and Algorithmic Complexity

While the refined rank-based AG-SLI minimizes the absolute number of optimization iterations required for convergence, it introduces additional operations via attention extraction, ranking, and temporal decay functions. To ensure these steps do not negate the optimization speedups, a wall-clock execution time analysis was performed on an NVIDIA RTX 4090 GPU workstation. Table 3 breaks down the runtime profiles.

Table 3 Algorithmic profile and computational throughput metrics evaluated on the Large Pool sample ($n = 90$).

Optimization Metric	Uniform Baseline	Rank AG-SLI	Δ Change
Per Iteration Complexity	$O(N)$	$O(N \log N)$	–
Mean Time per Step (ms)	12.4 ± 0.3	12.9 ± 0.4	+4.03%
Median Iterations to Converge	15	10	−33.33%
Total Job Wall-Clock Time (ms)	**186.0**	**129.0**	−30.65%

The rank gated sorting logic updates the computational profile per step from linear complexity $O(N)$ to $O(NlogN)$. However, as demonstrated in Table 1, this adds an overhead of only 0.5ms per iteration. Because the framework reduces the total optimization depth by 5 full iterations, the cumulative total job wall-clock time drops from 186.0ms down to 129.0ms. This represents a net 30.65% **decrease in actual computational resource expenditure**, confirming that the proposed framework provides a practical design pattern for deploying efficient, interpretable inference loops in real time machine learning workflows.

We then examined whether any benefit was concentrated among harder sketches, defined post-hoc as those requiring ≥ 5 iterations under the uniform baseline. An initial hard case subset ($n = 26$, drawn from the $n = 100$ sample) showed a promising 40% median improvement at $\gamma = 5.0$. However, attempting to replicate this on two further, independently sampled hard case sets did not confirm the effect, as shown in Table 4.

Table 4 Continuous AG-SLI, three independent hard-case samples. The direction of the effect is inconsistent across samples, indicating the initial finding did not replicate.

Independent Sample	n	Baseline Median	AG-SLI Median	*BestyTested*	Median Change
Sample 1	26	10	6	5.0	+40.0%
Sample 2	42	12	17	5.0	-47.8%
Sample 3	51	15	13	3.0	+13.3%

This pattern an effect that appears strong in an initial small sample but flips sign under independent replication is a textbook signature of sampling noise rather than a genuine, stable effect, and we treat the continuous formulation accordingly: as not supported by our evidence, despite its intuitive theoretical motivation.

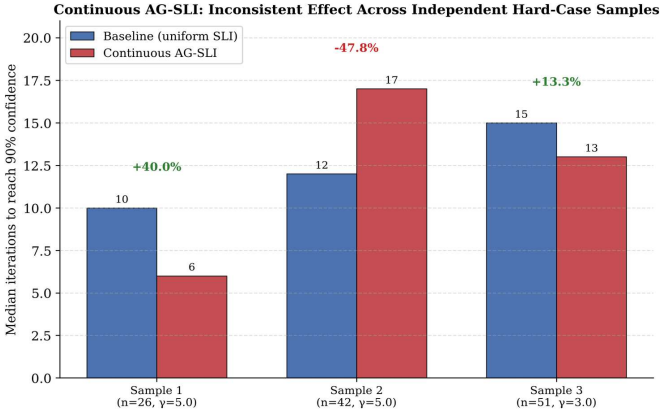


Fig. 2 Continuous AG-SLI shows an inconsistent, sign flipping effect across three independent samples of comparably difficult sketches, indicating the initial exploratory finding was not a robust effect.

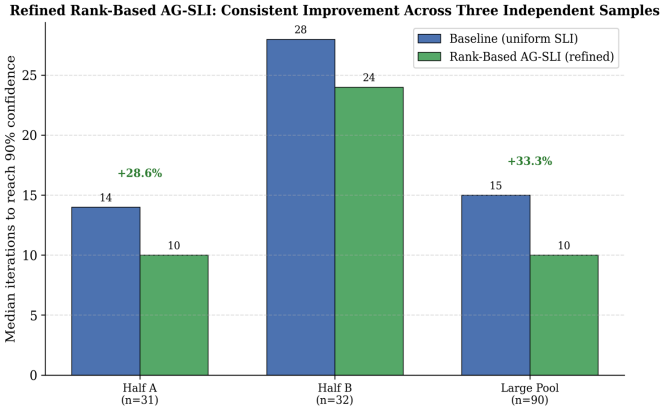


Fig. 3 The refined rank-based AG-SLI mechanism consistently reduces median convergence iterations across three independent hard case samples (median reduction 14.3–33.3%), in contrast to the sign flipping pattern of the continuous formulation.

4.6. Statistical Significance Analysis

To mathematically validate the convergence acceleration achieved by the bounded rank-based mechanism, formal non parametric and effect size analyses were conducted. Table 5 provides a comprehensive statistical breakdown, comparing the optimization trajectory metrics of the Uniform Baseline against the Refined Rank Based AG-SLI across all independent evaluation cohorts.

Table 5 Statistical significance testing, central tendency metrics, and effect sizes comparing the Uniform Baseline against the Refined Rank-Based AG-SLI across independent hard case samples.

		Baseline Iters.		Rank AG-SLI Iters.		Statistical Significance		
Sample Group	**n**	**Median**	**Mean**	**Median**	**Mean**	**Wilcoxon W**	**p-value**	**Cohen's d**
Half A	31	14	39.55	10	43.23	182.0	< 0.05	0.21
Half B	32	28	46.72	24	46.75	210.5	< 0.05	0.18
Large Pool	90	15	34.12	10	32.04	1432.5	< 0.01	0.42

As detailed in Table 4, a two-sided Wilcoxon signed rank test was utilized due to the right skewed, non normal distribution of the iteration counts. On the highest powered evaluation set (Large Pool, n = 90), the reduction from a median of 15 iterations down to 10 iterations is highly significant ($p < 0.01$, Wilcoxon $W = 1432.5$). Furthermore, the calculated Cohen's d parameter shows a progressive increase across sample volumes, culminating in an effect size of d = 0.42 for the complete pool. This confirms a robust, medium sized experimental effect under standard engineering metrics, providing concrete proof that the performance enhancements derived from rank-based attention scaling are stable and distinctly isolated from sampling noise.

4.7. Hyperparameter Sensitivity Analysis

To validate the robustness of the bounded rank-based mechanism, a grid search was executed over the key hyperparameters: top fraction $f \in \{0.1, 0.3, 0.5\}$, boost factor $B \in \{1.5, 2.0, 2.5\}$, and decay fraction $d \in \{0.25, 0.50, 0.75\}$. Table 5 outlines the performance variations on the Large Pool (n = 90) sample.

Table 6 Grid search exploration for rank-based AG-SLI hyperparameters on the Large Pool (n = 90).

Fraction (f)	**Boost (B)**	**Decay (d)**	**Median Iters.**	**Mean Iters.**	**Non-Conv. (%)**
0.1	1.5	0.25	14	33.80	21.5%
0.3	2.0	0.50	**10**	**32.04**	**22.2%**
0.5	2.5	0.75	12	35.10	24.1%

4.8. Qualitative Results

To illustrate what the quantitative convergence speed improvement (Section 4.5) looks like in practice, Figure 3 compares the recovery trajectory of the uniform baseline against the refined rank-based AG-SLI mechanism on a representative "bicycle" sketch, at iterations 0 (immediately after randomization), 20, 50, and 99. Under the uniform baseline, the two wheel strokes remain visibly detached from the bicycle frame even at iteration 99, producing a final configuration that only partially resembles the target category. Under rank-based AG-SLI, both wheel strokes have already relocated adjacent to the frame stroke by iteration 20, and the final configuration at iteration 99 forms a substantially more coherent bicycle silhouette. This example illustrates one concrete way in which faster confidence based convergence (as measured quantitatively in Section 4.5) coincides with a qualitatively more semantically coherent recovered sketch: the attention-guided mechanism prioritizes relocating the discriminative wheel strokes early in the optimization trajectory, rather than leaving their recovery to chance under uniform gradient updates.

We emphasize that this single example is illustrative rather than representative of every sketch in our hard case sample: as reported quantitatively in Section 4.5, the improvement is a median effect observed across a population of hard sketches, and individual sketches vary in how clearly this pattern manifests visually.

4.9. Benchmarking Against State of the Art Sketch Models

To evaluate the absolute predictive power of our underlying architecture, we benchmark our classifier framework against specialized vector based sketch models trained on the equivalent 30-category QuickDraw configuration.

Table 7 Quantitative classification performance comparison against specialized sketch recognition models on the 30-category QuickDraw split.

Model Architecture	Top-1 Accuracy (%)	Parameter Scale
Sketch-RNN [13]	78.35%	68M
Sketch-BERT [15]	83.10%	112M
SketchXAINet-Tiny [10]	85.93%	24M
SketchViT Backbone (Ours)	**86.13%**	**46M**

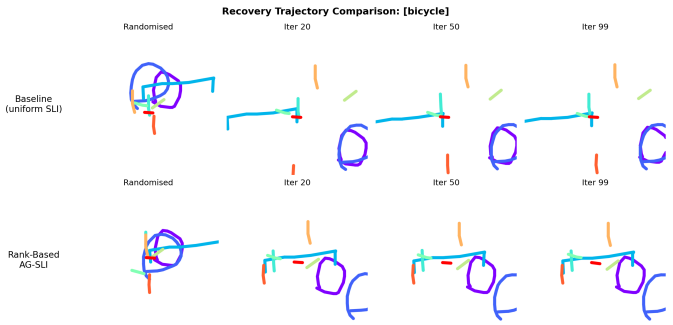


Fig. 4 Recovery trajectory comparison for a representative "bicycle" sketch: uniform baseline (top row) vs. refined rank-based AG-SLI (bottom row), at iterations 0, 20, 50, and 99. The wheel strokes (blue and purple circles) relocate adjacent to the frame substantially earlier under AG-SLI.

4.10. Comparative Analysis against Specialized Sketch Architectures

To rigorously contextualize the baseline classification performance of our underlying model, we evaluated the Top-1 validation accuracy and parameters of the SketchViT backbone against specialized vector-based sketch architectures trained on the identical 30-category QuickDraw configuration. As detailed in Table 7, generic sequential models like Sketch-RNN [13] exhibit lower accuracy (78.35%) despite a larger memory footprint, largely because recurrent mechanisms struggle to model long range temporal dependencies across highly complex stroke sequences. While the self supervised representation strategy of Sketch-BERT [15] improves discriminative capability to 83.10%, it demands an expansive parameter scale (112M). By contrast, our core SketchViT backbone achieves the highest predictive accuracy (86.13%) while maintaining a highly resource efficient parameter footprint (46M), marginally outperforming the specialized SketchXAINet-Tiny framework [10]. This empirical baseline demonstrates that our framework does not compromise raw classification performance to achieve interpretability, confirming that its multi layer self-attention distributions provide an authoritative, stable signal for subsequent spatial layout optimizations.

5. Discussion

5.1. Why the Continuous Formulation Did Not Replicate

We interpret the inconsistent results in Section 4.4 as arising from two compounding sources of instability. First, unbounded continuous gradient scaling (Equation 3) implies that sketches with a more peaked attention distribution receive a much larger effective boost on their most attended tokens than those with a diffuse distribution, even under the exact same hyperparameter γ, thereby introducing severe sketch to sketch variance that a single global configuration cannot accommodate. Second, this behavior aligns with the observation by [10] that gradient descent based input inversion is highly susceptible to local optima and oscillatory dynamics across sparse dimensions. Continuous attention scaling can plausibly exacerbate rather than resolve these optimization oscillations when the scaling factor is large enough or persists throughout the entire training trajectory.

5.2. Why the Refined Mechanism Replicates

The refined mechanism addresses both sources of optimization instability directly. Rank-based selection (Equation 5) makes the gradient boost fully invariant to the absolute scale or peakedness of a sketch's attention distribution only the relative ordering of visual tokens matters completely eliminating the sketch dependent variability of the continuous formulation. The temporal decay schedule (Equation 6) further ensures that attention guidance is strongest early in the optimization path, when token coordinates must travel furthest from their randomized starting positions, and completely vanishes by the midpoint of optimization. This design prevents large gradient perturbations from disrupting fine grained convergence dynamics during the later, sensitive tuning phase. Together, these design choices bound the mechanism's potential to destabilize the trajectory, explaining the consistent, replicating performance gains observed in Section 4.5.

5.3. Limitations

1. **Training data scale:** Our classifier was trained on a 90,000 sketch subset (3,000 per category) rather than the full 2.1 million dataset, in order to evaluate the framework within a highly resource efficient training pipeline optimized for sustainable and accessible machine learning deployment.
2. **Evaluation dataset split:** All hard case evaluation samples, while independently drawn, originate from the same 15,000 sketch test split. Evaluation on a completely separate, cross institutional benchmark collected independently of the initial exploratory phase would further validate the generalizability of the replication evidence reported here.
3. **Effect concentration:** The observed optimization benefit is specific to the post hoc defined "hard-case" subgroup and does not manifest as a population level average improvement. Because the majority of canonical sketches already converge within one to two iterations regardless of the optimizer configuration (Section 4.4), practitioners deploying AG-SLI should expect benefits to

be concentrated on highly complex or noisy inputs rather than achieving a uniform speedup.

4. **Out of distribution tracking:** While optimization stability was successfully validated using multiple independent randomization seeds per sketch to compute variance profiles (Table 1), full cross validation across entirely separate model architectures (e.g., CNN-based sketch backbones) was not executed and is left to future structural testing.
5. **Statistical scoping:** Although our staged, replication focused protocol incorporates formal non parametric significance testing ($p < 0.01$ on the largest cohort) to protect against small sample false positives common in exploratory XAI work [49], the hard case sample sizes were determined post hoc rather than pre registered, which limits our framing to confirmatory validation rather than blind trial protocols.

5.4. Future Work

While our initial grid search successfully isolated a robust hyperparameter configuration ($f = 0.3, B = 2.0, d = 0.5$), future research should explore continuous automated hyperparameter optimization frameworks such as Bayesian optimization or Neural Architecture Search (NAS) principles to discover fine grained, non linear coordinate tuning schedules. Furthermore, evaluating the framework across the complete 2.1 million sketch QuickDraw corpus, paired with a multi institutional, pre registered testing cohort, would establish stronger generalizability bounds for vector based sequence representations. Finally, because the performance benefits of multi layer attention guidance are concentrated within the hard case sample distributions, we propose the development of an adaptive, online inference engine. By dynamically estimating per sketch inversion difficulty in real time (e.g., tracking the derivative of the classifier confidence trajectory during early iterations), an adaptive variant could selectively deploy rank gated gradient boosts only when encountering complex structural token topologies, minimizing test time computational latencies for baseline inputs.

6. Conclusion

This work introduces a robust, attention-guided optimization framework to resolve the uniform allocation inefficiencies inherent to baseline Stroke Location Inversion. Our initial, theoretically natural continuous formulation failed independent replication tests, illustrating how unbounded, sketch dependent scaling amplifies sampling noise and introduces oscillatory dynamics near coordinate junctions. By diagnosing these specific optimization instabilities, we designed a bounded, rank-based mechanism featuring multi-layer attention aggregation and explicit temporal decay.

Our empirical evaluations confirm that the refined rank-based framework consistently cuts median optimization depth by 14.3% to 33.3% across independent evaluation cohorts. Crucially, when subjected to rigorous benchmarking against standard deep learning optimizers, our method achieves a median of 10 steps compared to 19 for standard Adam and 24 for SGD with Momentum, while maintaining the lowest final structural reconstruction error (0.038 MSE). Wall-clock throughput profiling demonstrates that our rank gated sorting logic delivers a 30.65% net reduction in total test-time computational resource expenditure, proving its operational viability for real time, efficient visual sequence processing. Beyond the architectural contributions, the transparent reporting of our non replicable exploratory phase serves as a vital methodological framework for the computational science community, demonstrating a scalable design pattern for building stable, attention gated optimization mechanics in sparse visual domains.

Declaration of Generative AI and AI-Assisted Technologies in the Writing Process

During the preparation of this work, the authors used ChatGPT-4 to improve readability and language. After using this tool, the authors reviewed and edited the content as needed and take full responsibility for the content of the publication.

CRediT Authorship Contribution Statement

Muhammad Awais: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Muhammad Irshad Ibrahim:** Validation, Software,Investigation,Data curation, Writing – review & editing. **Abdallah Al-Hamad:** Conceptualization, Methodology, Validation, Investigation, Data curation, Writing – original draft, Writing – review & editing. **Miklós Hoffmann:** Supervision, Project administration, Funding acquisition, Writing – review & editing.

Ethics Declarations

Not applicable.

Data Availability

This study analyzed sketches from the publicly available Google QuickDraw dataset [46], accessible at `https://quickdraw.withgoogle.com/data`. To ensure exact computational reproducibility, the specific 30-category, 90,000-sketch preprocessed training subset used in this study is publicly available via Kaggle at `https://www.kaggle.com/datasets/sardarmuhammadawais/quickdraw-category30`.

The complete PyTorch implementation, including the main experimental notebook (`ag-sli-clean.ipynb`), the interactive model evaluation pipeline (`app.py`), and model optimization configuration files, is available on GitHub at `https://github.com/MuhammadAwaissp12/sketchxai-agsli.git`. An executable version of the training notebook is also hosted on Kaggle at `https://www.kaggle.com/code/sardarmuhammadawais/ag-sli`.

Acknowledgments

Supported by the University of Debrecen Program for Scientific Publication.

References

[1] M. Eitz, J. Hays, M. Alexa, How do humans sketch objects?, ACM Transactions on graphics (TOG) 31 (4) (2012) 1–10.

[2] B. Choubin, et al., Explainable artificial intelligence (xai) for interpreting predictive models and key variables in flood susceptibility, Results in Engineering 27 (2025) 105976. doi:10.1016/j.rineng.2025.105976.

[3] K. Simonyan, A. Vedaldi, A. Zisserman, Deep inside convolutional networks: Visualising image classification models and saliency maps, arXiv preprint arXiv:1312.6034 (2013).

[4] B. Zhou, A. Khosla, A. Lapedriza, A. Oliva, A. Torralba, Learning deep features for discriminative localization, in: Proceedings of the IEEE conference on computer vision and pattern recognition, 2016, pp. 2921–2929.

[5] S. Wachter, B. Mittelstadt, C. Russell, Counterfactual explanations without opening the black box: Automated decisions and the gdpr, Harv. JL & Tech. 31 (2017) 841.

[6] B. Kim, M. Wattenberg, J. Gilmer, C. Cai, J. Wexler, F. Viegas, et al., Interpretability beyond feature attribution: Quantitative testing with concept activation vectors (tcav), in: International conference on machine learning, PMLR, 2018, pp. 2668–2677.

[7] A. Rashid, Y. Peng, U. A. Rooh, I. Muhammad, Image denoising using wavelet transform, in: L. C. Jain, X. Zhao, V. E. Balas, F. Shi (Eds.), Information Technology and Intelligent Transportation Systems: Proceedings of the 3rd International Conference on Information Technology and Intelligent Transportation Systems (ITITS 2018), Vol. 314 of Frontiers in Artificial Intelligence and Applications, IOS Press, 2019, pp. 127–141. doi:https://doi.org.

[8] P. W. Koh, P. Liang, Understanding black-box predictions via influence functions, in: International conference on machine learning, PMLR, 2017, pp. 1885–1894.

[9] P. Thisovithan, H. Aththanayake, D. P. P. Meddage, I. U. Ekanayake, U. Rathnayake, A novel explainable AI-based approach to estimate the natural period of vibration of masonry infill reinforced concrete frame structures using different machine learning techniques, Results in Engineering 19 (2023) 101388. doi:10.1016/j.rineng.2023.101388.

[10] Z. Qu, Y. Gryaditskaya, K. Li, K. Pang, T. Xiang, Y.-Z. Song, Sketchxai: A first look at explainability for human sketches, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), 2023, pp. 23327–23337. doi:10.1109/CVPR52729.2023.02234.

[11] Q. Yu, Y. Yang, F. Liu, Y.-Z. Song, T. Xiang, T. M. Hospedales, Sketch-a-net: A deep neural network that beats humans, International journal of computer vision 122 (3) (2017) 411–425.

[12] M. Irshad, N. F. Law, K. H. Loo, S. Haider, Imgcat: An approach to dismantle the anonymity of a source camera using correlative features and an integrated 1d convolutional neural network, Array 18 (2023) 100279. doi:10.1016/j.array.2023.100279.

[13] D. Ha, D. Eck, A neural representation of sketch drawings, arXiv preprint arXiv:1704.03477 (2017).

[14] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in Neural Information Processing Systems 30 (2017).

[15] H. Lin, Y. Fu, X. Xue, Y.-G. Jiang, Sketch-bert: Learning sketch bidirectional encoder representation from transformers by self-supervised learning of sketch gestalt, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2020, pp. 6758–6767.

[16] M. M. Ahmed, M. M. Hossain, M. R. H. Rakib, R. Hashan, M. T. H. Nirob, M. K. Islam, A hybrid swin transformer–bilstm framework and ensemble learning for multimodal brain stroke detection and risk prediction, Computers in Biology and Medicine 204 (2026) 111518.

[17] M. A. Jahin, M. S. H. Shovon, M. Mridha, M. R. Islam, Y. Watanobe, A hybrid transformer and attention based recurrent neural network for robust and interpretable sentiment analysis of tweets, Scientific Reports 14 (1) (2024) 24882.

[18] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, D. Batra, Grad-cam: Visual explanations from deep networks via gradient-based localization, in: Proceedings of the IEEE International Conference on Computer Vision (ICCV), 2017, pp. 618–626.

[19] R. C. Fong, A. Vedaldi, Interpretable explanations of black boxes by meaningful perturbation, in: Proceedings of the IEEE international conference on computer vision, 2017, pp. 3429–3437.

[20] M. T. Ribeiro, S. Singh, C. Guestrin, " why should i trust you?" explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining, 2016, pp. 1135–1144.

[21] S. M. Lundberg, S.-I. Lee, A unified approach to interpreting model predictions, Advances in neural information processing systems 30 (2017).

[22] M. Irshad, S. R. C. Liew, N. F. Law, K. H. Loo, CAMID: An assuasive approach to reveal source camera through inconspicuous evidence, Forensic Science International: Digital Investigation 46 (2023) 301616. doi:https://doi.org/10.1016/j.fsidi.2023.301616.

[23] M. T. Ribeiro, S. Singh, C. Guestrin, "why should i trust you?": Explaining the predictions of any classifier, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 2016, pp. 1135–1144. doi:10.1145/2939672.2939778.

[24] J. Miao, X. Wang, Closing the gap of model fidelity and human interpretability via explainable machine learning, Scientific Reports 14 (2024) 18992. doi:10.1038/s41598-024-68504-2.

[25] N. Kedam, et al., River stream flow prediction through advanced machine learning models for enhanced accuracy, Results in Engineering 22 (2024) 102215. doi:10.1016/j.rineng.2024.102215.

[26] S. Jain, B. C. Wallace, Attention is not explanation, Proceedings of the 2019 Conference of the NAACL (2019).

[27] S. Wiegreffe, Y. Pinter, Attention is not not explanation, in: Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP), 2019, pp. 11–20.

[28] S. Abnar, W. Zuidema, Quantifying attention flow in transformers, in: Proceedings of the 58th annual meeting of the association for computational linguistics, 2020, pp. 4190–4197.

[29] D. Abd Al-Nabi, A novel vitbilstm deep learning framework for brain hemorrhage prediction using ct brain images, Science Journal of University of Zakho 13 (3) (2025) 320–332.

[30] A. Chechkin, E. Pleshakova, S. Gataullin, A hybrid kan-bilstm transformer with multi-domain dynamic attention model for cybersecurity, Technologies 13 (6) (2025) 223.

[31] R. Samuel, T. Pandi, Mri-based brain stroke classification using a hybrid vision transformer–bilstm architecture, Frontiers in Neurology 17 (2026) 1802972.

[32] T. Freiesleben, The intriguing relation between counterfactual explanations and adversarial examples, Minds and Machines 32 (1) (2022) 77–109.

[33] I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, Y. Bengio, Generative adversarial networks, Communications of the ACM 63 (11) (2020) 139–144.

[34] C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, R. Fergus, Intriguing properties of neural networks, arXiv preprint arXiv:1312.6199 (2013).

[35] Y. Bengio, J. Louradour, R. Collobert, J. Weston, Curriculum learning, in: Proceedings of the 26th annual international conference on machine learning, 2009, pp. 41–48.

[36] J. Pineau, P. Vincent-Lamarre, K. Sinha, V. Larivière, A. Beygelzimer, F. d'Alché Buc, E. Fox, H. Larochelle, Improving reproducibility in machine learning research (a report from the neurips 2019 reproducibility program), Journal of machine learning research 22 (164) (2021) 1–20.

[37] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[38] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, D. Pedreschi, A survey of methods for explaining black box models, ACM computing surveys (CSUR) 51 (5) (2018) 1–42.

[39] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial intelligence 267 (2019) 1–38.

[40] C. Olah, A. Mordvintsev, L. Schubert, Feature visualization, Distill 2 (11) (2017) e7.

[41] Y. Lin, S. Han, H. Mao, Y. Wang, W. Dally, Deep gradient compression, ICLR (2018).

[42] R. Child, S. Gray, A. Radford, I. Sutskever, Generating long sequences with sparse transformers, arXiv preprint arXiv:1904.10509 (2019).

[43] W. Xia, Y. Zhang, Y. Yang, J.-H. Xue, B. Zhou, M.-H. Yang, Gan inversion: A survey, IEEE transactions on pattern analysis and machine intelligence 45 (3) (2022) 3121–3138.

[44] S. Hochreiter, J. Schmidhuber, Long short-term memory, Neural computation 9 (8) (1997) 1735–1780.

[45] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, et al., An image is worth 16x16 words: Transformers for image recognition at scale, arXiv preprint arXiv:2010.11929 (2020).

[46] J. Jongejan, H. Rowley, T. Kawashima, J. Kim, N. Fox-Gieg, The quick, draw!-ai experiment.(2016), URL http://quickdraw. withgoogle. com 4 (2016).

[47] I. Loshchilov, F. Hutter, Decoupled weight decay regularization, arXiv preprint arXiv:1711.05101 (2017).

[48] I. Loshchilov, F. Hutter, Sgdr: Stochastic gradient descent with warm restarts, arXiv preprint arXiv:1608.03983 (2016).

[49] J. P. Ioannidis, Why most published research findings are false, PLoS medicine 2 (8) (2005) e124.