Published by :
https://www.ijert.org/
An International Peer-Reviewed Journal

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Why AI Gets Black Hair Wrong: Documented Failure Modes and Architectural Fixes in a Production Try-On System

Dr. Isi Idemudia[1] Christelle Mombo-Zigah[2]
[1] *Chief Technology Officer* [2] *Founder & Chief Executive Officer*
StyleMyCrown
San jose California · USA

Abstract - Generative image-editing models are increasingly used for virtual hairstyle try-on, yet their behavior on Black hair textures and on darker-skinned subjects is poorly characterized. We report on the design and operation of a production try-on system serving a global user base, in which every generation is validated by an instrumented, multi-gate pipeline that logs per-candidate diagnostics. From this telemetry we derive a reproducible taxonomy of five failure modes observed in composite (two-image) generative editing: catalog echo (reproduction of the reference image), no-op (unchanged input returned), identity transfer / prior drift (substitution of the subject with a different, frequently lighter-skinned person), timid-edit under-transformation, and capture-condition segmentation failure. We show that several of these failures are correlated with subject skin tone and with the statistical improbability, under the model's learned prior, of a given (face, hairstyle) pairing. We then describe the architectural response: a masked-inpainting path that makes identity preservation a physical property of the pipeline rather than a prompt-level request, a layered validation gate stack in which no single instrument can convict or clear a candidate alone, and a set of refusal semantics that convert unsafe generations into credit-preserving refusals rather than harmful deliveries. We argue that for identity-critical, culturally-specific generative applications, correctness must be enforced architecturally and measured continuously, and we offer our incident-driven gate ledger as a template.

Keywords - ***generative image editing; algorithmic bias; identity preservation; Black hair; virtual try-on; diffusion models; production ML systems; fairness; inpainting***

1. INTRODUCTION

Virtual try-on applications let a user upload a photograph and preview themselves in a different hairstyle. The underlying technology, text-guided image editing built on diffusion models [3] has advanced rapidly, but its evaluation has centered on aggregate visual quality rather than on identity preservation, and its training distributions underrepresent Black hairstyles and darker-skinned faces [2], [6]. For an application whose explicit purpose is to serve Black users and to render culturally-specific styles such as box braids, cornrows, locs, and bantu knots, this gap is not a peripheral concern; it is the central engineering problem.

This paper reports on a production system, StyleMyCrown, that performs hairstyle try-on for a global community with a focus on Black hair textures. Rather than treating the generative model as a trustworthy black box, the system wraps every generation in an instrumented validation pipeline that records, for each candidate image, a structured diagnostic object: a facial-embedding distance, a vision-language-model (VLM) quality score for both face preservation and hairstyle accuracy, perceptual-hash comparisons against both the user photograph and the catalog reference, and the resulting gate decision. Because these diagnostics are logged for every transform, the system doubles as a measurement instrument for its own failure modes.

We make three contributions. First, we present a reproducible **taxonomy of five failure modes** observed in composite generative hairstyle editing, each defined by a distinct signature in the logged diagnostics. Second, we present **evidence that several of these failures are correlated with skin tone** and with the improbability of a (face, hairstyle) pairing under the model's learned prior a mechanism we distinguish from simple image-quality degradation. Third, we describe the **architectural response**: a masked-inpainting pipeline that enforces identity preservation physically, a layered gate stack with explicit fail-open and fail-closed semantics, and an incident-driven governance discipline in which every validation rule is traceable to a specific production failure.

2. SYSTEM OVERVIEW

The system accepts a user photograph and a selected catalog style, and returns either a transformed image or a credit-preserving refusal. Generation is performed by one of two paths, and validation is shared across both.

2.1 Generation paths

The **composite path** presents the generative editor with the user photograph and the catalog reference image and requests that the reference hairstyle be applied to the user. This path is general, it serves any style but, as we document, it is where identity failures concentrate.

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

The **masked-inpainting path** first computes a semantic segmentation mask that protects the facial region, then permits generation only within the unprotected (hair) region. Here identity preservation is not requested from the model; it is a physical property of the pipeline, because the model is architecturally prevented from altering the protected pixels. This distinction—identity as a request versus identity as a guarantee, is the paper's central architectural claim.

2.2 The validation gate stack

Regardless of generation path, each candidate passes through a shared validation stack. Multiple independent instruments are consulted: a facial-recognition model yields a Euclidean embedding distance between the user and the candidate; a VLM evaluator scores face preservation and hairstyle accuracy on a defined rubric; and perceptual hashing measures structural similarity of the candidate to both the user photo and the catalog reference. A selection module then resolves these signals into one of three outcomes; deliver, deliver-with-low-confidence, or refuse-and-refund, under the decision rules of Section 5. A design invariant, discussed throughout, is that no single instrument may unilaterally convict or clear a candidate; conviction requires corroboration, and a single reliable contrary signal can veto delivery.

Figure 3. System architecture and outcome resolution

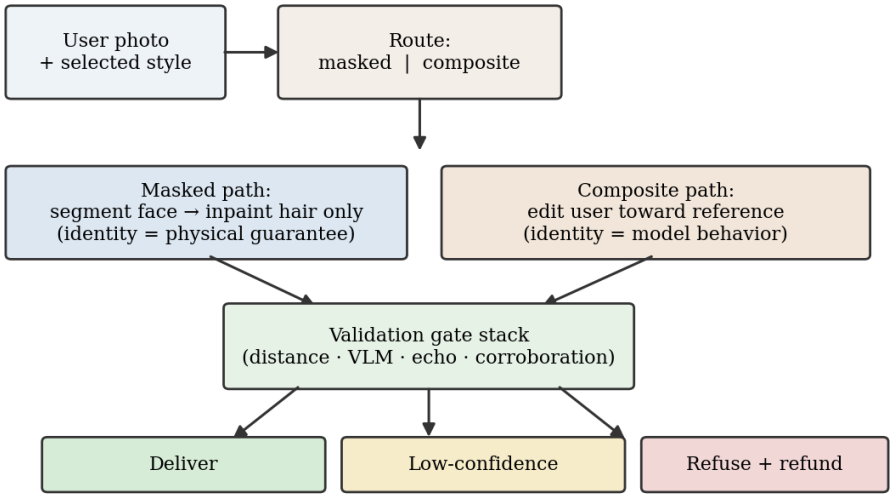
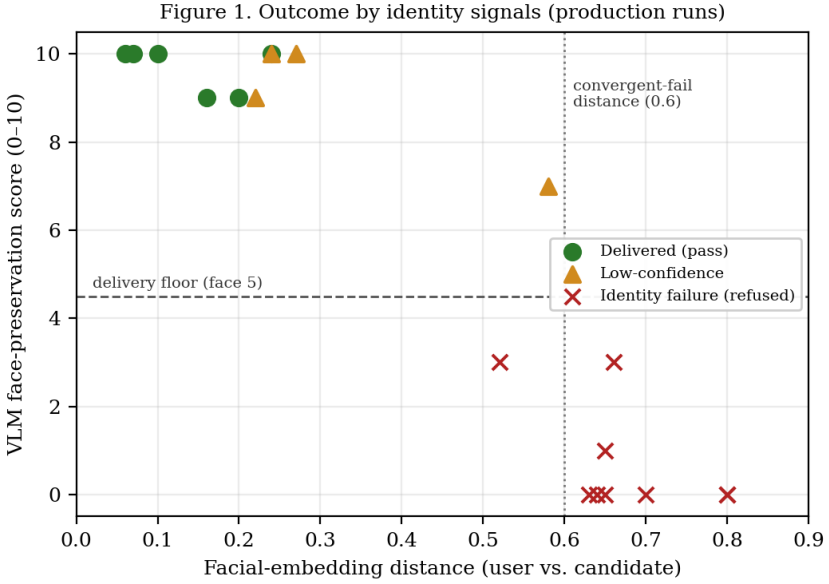


Figure 3. Two generation paths feed a shared validation gate stack that resolves every candidate into one of three outcomes. On the masked path, identity preservation is a physical property; on the composite path, it is a hoped-for behavior of the model.

3. A TAXONOMY OF FAILURE MODES

We define each failure mode by its signature in the logged diagnostics, so that it is detectable and reproducible rather than anecdotal. Table 1 summarizes the taxonomy; the subsections elaborate. Figure 1 plots representative production runs in the two-dimensional space of the identity instruments, showing how the outcomes separate.

Figure 1. Outcome by identity signals (production runs)



International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

Figure 1. Each point is a production candidate positioned by its two identity signals. Delivered results cluster at low embedding distance and high face score; identity failures (refused) cluster at high distance and near-zero face score. The delivery floor and convergent-failure boundary are shown.

Table 1. Failure modes in composite generative hairstyle editing.

Mode	Diagnostic signature	Description
Catalog echo	Low hash distance to catalog reference; high hairstyle score; low face score	The output reproduces the reference image rather than editing the user; the reference model's face appears in the result
No-op	Very low hash distance to user photo; low hairstyle score; high face score	The user's own photo is returned essentially unchanged; the requested style is not applied
Identity transfer / prior drift	High embedding distance; high hairstyle score; face score at or near zero	The subject is replaced by a different, frequently lighter-skinned, person who fits the model's prior for the requested style and some occasions replaced with dark-skinned.
Timid edit	Moderate face score; low hairstyle score; user-echo flag	Identity is preserved but the transformation is under-committed, especially for large or high-volume styles
Segmentation failure	Near-total paintable mask; large post-hoc embedding distance on the masked path	On backlit or low-contrast photographs of darker-skinned faces, face segmentation fails, removing the identity guarantee

3.1 Catalog echo

In catalog echo the generative editor, presented with two images, returns something close to a copy of the reference rather than an edit of the user. The signature is unambiguous: the perceptual-hash distance from the candidate to the catalog reference falls below threshold while the distance to the user photo remains large; the VLM scores hairstyle accuracy very high (the reference hair is reproduced faithfully) and face preservation at or near zero (it is the reference model's face, not the user's). We observed candidates at catalog hash distances of 98–120 bits on a 1024-bit hash, i.e. near-structural copies.

3.2 No-op

The inverse failure returns the user's photograph essentially unchanged. Here the hash distance to the user photo is very small and the VLM scores hairstyle accuracy at or near zero while face preservation is high. Delivering a no-op is delivering nothing; because a plausible face score can accompany it, a naive face-only gate will pass it. This mode motivated a dedicated detector that requires agreement between a low hairstyle score and either near-zero user-distance or a user-echo hash flag before refusing.

3.3 Identity transfer and prior drift

The most consequential mode replaces the subject with a different person. Two instruments agree: the facial-embedding distance is large and the VLM scores face preservation at zero ("different person"), while hairstyle accuracy scores high. We distinguish two sub-cases. In **reference-anchored transfer**, the substituted face is borrowed from the catalog reference model; in **prior drift**, the model synthesizes an entirely new person, frequently lighter-skinned than the user, whose appearance better matches its learned expectation of who wears the requested style. The distinction is testable: reference-anchored transfer produces intermediate catalog-hash distances, whereas prior drift does not resemble the catalog structurally at all. Section 4 develops the evidence that this mode is skin-tone-correlated.

3.4 Timid edit

A subtler failure preserves identity but under-transforms. It appears most often on large, high-volume styles that require painting hair well beyond the original silhouette. Face preservation scores well and embedding distance is small, but hairstyle accuracy lands in the "same category, different execution" band. In our data the composite path was bimodal on high-transformation styles: on a given run it either copied the reference (Section 3.1/3.3) or produced a timid edit, but rarely landed the requested transformation on the correct person, an observation that, as we argue in Section 6, points to an architectural rather than a prompt-level cause.

3.5 Capture-condition segmentation failure

The masked path's identity guarantee depends on successfully locating the face. On backlit or low-contrast photographs of darker-skinned subjects, ordinary outdoor conditions for many users—the segmentation model failed to detect the face, producing a mask whose protected region was near-empty. In one measured incident the protected fraction was effectively zero, and the model repainted the entire image. The lesson generalizes: an architectural guarantee inherits the weakest link in its dependency chain, and here that link (segmentation) shares the same demographic weakness the whole system exists to overcome.

4. SKIN TONE AND THE LEARNED PRIOR

The failures of Sections 3.1 and 3.3 are not uniformly distributed across users. On identical styles, identical pipeline, and comparable photograph quality, transforms that succeeded for lighter-skinned subjects failed by identity transfer for darker-skinned subjects, including subjects of different ethnic backgrounds who share deep skin tones. Because the controlled variable in these

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

paired observations was the subject's face rather than photograph quality, style, or pipeline, we attribute the difference to the generative model's learned prior over (face, hairstyle) pairings.

The mechanism is as follows. A generative editor trained on internet-scale imagery [6], [7] absorbs the empirical correlations of that corpus, including which faces co-occur with which hairstyles. When a requested (face, hairstyle) pairing is improbable under this prior because the corpus rarely depicts that combination, the model tends to resolve the tension by altering the more weakly-constrained variable. When the instruction constrains the hairstyle strongly, the face is what gives way. The result is a technically excellent hairstyle rendered on a substituted, prior-consistent, and frequently lighter-skinned face. In effect, the model encodes an assumption about who wears a given style, and edits the user to fit it.

Two further observations sharpen the account. First, the failure is **input-dependent, not merely style-dependent**: a style that transformed correctly for one user underwent identity transfer for another on a later run, so the property cannot be predicted from the style alone. Second, the failure is **stochastic**: repeated runs of the same (photo, style) pair did not fail identically, consistent with the bimodal behavior of Section 3.4. Both observations bear directly on mitigation, since neither a per-style blocklist nor a single deterministic patch can fully address an input-dependent, stochastic failure.

We also note a measurement hazard. The two identity instruments do not agree uniformly across skin tones: on darker-skinned subjects we observed compressed facial-embedding distances, smaller separations between genuinely different people, while the VLM evaluator continued to score such substitutions as failures. This is consistent with the documented tendency of facial-analysis systems to perform worst on darker-skinned subjects [1]. A pipeline that trusted the embedding distance alone would therefore be least protective precisely for the users most exposed to the underlying failure. This is the empirical basis for the corroboration requirement of Section 5, and it cautions that a fairness evaluation must characterize the bias of its own evaluators rather than treat them as ground truth.

5. LAYERED VALIDATION AND REFUSAL SEMANTICS

The validation stack is designed around a single principle: for an identity-critical application, delivering a wrong face is a categorically worse outcome than refusing a possibly-acceptable one. The gates therefore enforce an asymmetry between the evidence required to deliver and the evidence required to refuse.

5.1 Corroboration and the no-single-instrument rule

No instrument convicts alone. A verified identity failure requires either a very large embedding distance, or the agreement of two independent instruments, for example, a near-zero VLM face score together with an elevated embedding distance ("convergent failure"), or a catalog-echo hash flag together with a near-zero face score. Symmetrically, perceptual-hash echo flags are treated as corroborating evidence only: because a high-fidelity edit can legitimately resemble the input in structure, a candidate that passes the full gate is delivered even if echo-flagged. This rule prevents both false refusals from an over-eager hash and false deliveries from a single miscalibrated signal.

5.2 The delivery floor

Independent of conviction, the system enforces a delivery floor: a candidate whose VLM face-preservation score falls below the rubric's "same person recognizable" threshold is never delivered, even as a low-confidence fallback, even absent a corroborating instrument. This rule was introduced after a candidate scored just above an earlier, lower floor was delivered as a blended, partially-substituted face. Calibrating the floor to the recognizability boundary of the rubric closed the gap between what the evaluator detected and what the system permitted.

5.3 Fail-open versus fail-closed

The stack distinguishes infrastructure failure from identity failure. When an instrument is genuinely unavailable; a model fails to load, an upload errors, a response fails to parse, the system fails **open**, delivering the best available candidate with a low-confidence flag rather than blaming the user or blanking the screen; a missing measurement is never silently treated as a score of zero. When identity failure is affirmatively verified, the system fails **closed**, refusing and refunding. This separation prevents transient infrastructure faults from manifesting as "we couldn't preserve your face" messages, which would both mislead the user and mask real outages.

5.4 Refusal as a first-class outcome

Because the gates catch identity failures reliably, the user-visible consequence of the biases in Section 4 is not a wrong face but a refusal. This is the correct safety behavior, yet it carries its own equity cost: refusals are distributed unevenly across skin tones, so the users the model serves least are refused most. We therefore treat refusal rate by subject characteristics as a monitored fairness metric, not merely an error rate, and we regard reducing it—rather than merely refusing safely—as the outstanding problem.

6. IDENTITY AS A PHYSICAL GUARANTEE

The recurring lesson across failure modes is that prompt-level instructions to preserve identity are unreliable, because the model may satisfy a strongly-constrained hairstyle request by sacrificing a weakly-constrained face. The masked-inpainting path removes

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

this degree of freedom. A semantic segmentation model [8] protects the facial region, and generation is permitted only in the complementary hair region; the protected pixels cannot be altered because the model never generates them. Identity preservation thereby becomes a property of the pipeline geometry rather than a hoped-for behavior of the model.

The empirical contrast is stark. On the masked path, embedding distances between input and output clustered near zero, and identity-transfer and prior-drift modes were absent by construction, because the face is not a region the model is free to reinterpret. The trade-off is that hairstyle fidelity on the masked path depends on the generator's competence within the hair region and, for highly-specific styles, benefits from style-specialized model adapters such as low-rank fine-tuning [4], [5]. The architecture thus separates two concerns that the composite path conflates: **identity**, which the mask guarantees for every style, and **style fidelity**, which can be improved per-style without ever placing identity at risk. Figure 2 shows the resulting separation in measured identity drift.

Figure 2. Identity drift by generation path

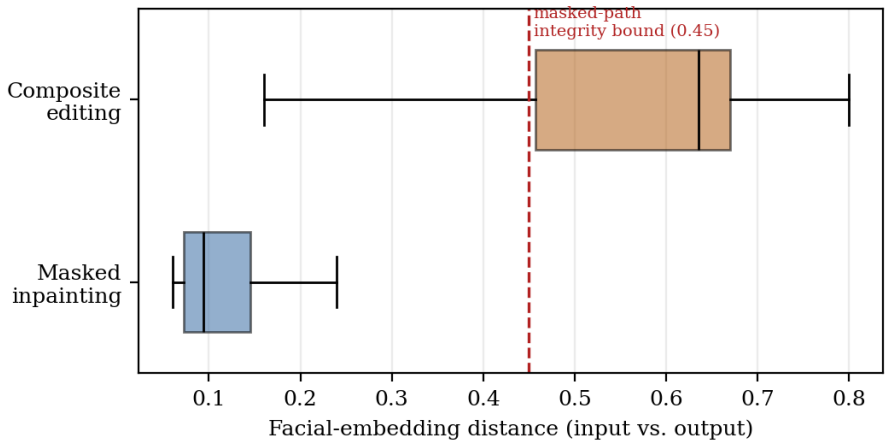


Figure 2. Input-to-output embedding distance by generation path. Masked-path outputs remain near the input identity; composite-path outputs spread toward and past the identity-failure region. The masked-path integrity bound flags any masked output whose distance indicates the mask did not hold.

Two caveats preserve honesty about the guarantee. First, as Section 3.5 shows, the guarantee is only as strong as the segmentation it rests on, and segmentation itself exhibits the demographic weakness the system exists to counter; hardening detection on real-world capture conditions is therefore a first-order safety task, not a quality nicety. Second, the masked path is incompatible with styles that must alter the protected region itself—wigs with lace fronts that cross the hairline—which must be routed differently. A guarantee with clearly-stated boundaries is more useful than an absolute claim that fails silently outside them.

7. INCIDENT-DRIVEN GOVERNANCE

Every rule in the validation stack is traceable to a specific production incident. We maintain this correspondence explicitly as a gate ledger (Table 2), in which each rule records the failure that motivated it. This discipline has two benefits. It prevents well-intentioned simplification: a maintainer who cannot see why a rule exists is apt to remove it, and a rule annotated with the incident it prevents resists that. And it turns the validation layer into a living specification of the system's known failure surface, which is precisely the artifact an audit or a fairness review requires.

Table 2. Representative gate ledger: each rule traces to an incident.

Validation rule	Motivating incident
Fail-open on infrastructure error	Silent credit loss when a validator outage rejected all candidates
Verified embedding-distance threshold	Delivery of visibly wrong faces
Convergent identity failure (two instruments)	Wrong faces at embedding distances just below the single-instrument threshold
Catalog-echo detection	Reference image delivered to a user in place of an edit
No-op / user-echo detection	Unchanged user photos delivered as "results"
Delivery floor at recognizability	A blended, partially-substituted face delivered as low-confidence
Masked-path distance invariant	Segmentation miss allowed the face to be repainted on a "protected" path
Degenerate-mask guard	Near-empty protective mask on a backlit dark-skinned photograph

International Journal of Engineering Research & Technology (IJERT)
ISSN: 2278-0181
Vol. 15 Issue 07 , July - 2026

8. DISCUSSION AND LIMITATIONS

Our evidence is observational, drawn from production telemetry rather than a controlled benchmark, and our subject counts in the paired skin-tone comparisons are small. We therefore frame the skin-tone results as a well-characterized mechanism supported by reproducible signatures, not as a quantified effect size; a controlled study across a stratified set of skin tones, styles, and capture conditions is the natural next step, and the diagnostic logging described here is designed to produce exactly that dataset. A second limitation is that our VLM evaluator is itself a potential source of bias; we have relied on the disagreement between it and the embedding instrument as a check, but a rigorous fairness evaluation must characterize each evaluator's own calibration across skin tones. Finally, our findings concern the specific models deployed during the study period; generative models change, and the failure modes, though we expect them to persist in kind—will shift in frequency and detail.

With those limits stated, the architectural conclusions are robust to the particular model in use. The superiority of a physical identity guarantee over a prompt-level request, the necessity of corroboration among independent instruments, the asymmetry between the evidence required to deliver and to refuse, and the value of an incident-traceable gate ledger are all properties of the system design rather than of any one generator.

9. CONCLUSION

AI gets Black hair wrong in specific, reproducible ways: it copies the reference, it returns the input unchanged, it under-commits the edit, and most seriously, it substitutes the subject with a prior-consistent, frequently lighter-skinned person and occasionally with darker-skinned person when the requested (face, hairstyle) pairing is improbable under its training distribution. These are not random defects but structured consequences of what the underlying models have learned. In a production system serving Black users, we found that the reliable remedy was not better prompting but better architecture: enforce identity as a physical property of the pipeline where possible, validate every generation with layered and corroborating instruments, prefer a credit-preserving refusal to a harmful delivery, and record every safeguard against the incident that motivated it. Making these systems correct, and keeping them correct as their models drift, requires treating measurement and refusal as first-class features rather than afterthoughts.

REFERENCES

[1] J. Buolamwini and T. Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," in Proc. 1st Conf. on Fairness, Accountability and Transparency (FAT*), PMLR, vol. 81, pp. 77–91, 2018.
[2] A. Birhane, V. U. Prabhu, and E. Kahembwe, "Multimodal datasets: misogyny, pornography, and malignant stereotypes," arXiv preprint arXiv:2110.01963, 2021.
[3] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, "High-Resolution Image Synthesis with Latent Diffusion Models," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 10684–10695, 2022.
[4] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman, "DreamBooth: Fine Tuning Text-to-Image Diffusion Models for Subject-Driven Generation," in Proc. IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 22500–22510, 2023.
[5] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-Rank Adaptation of Large Language Models," in Proc. Int. Conf. on Learning Representations (ICLR), 2022.
[6] C. Schuhmann, R. Beaumont, R. Vencu, et al., "LAION-5B: An Open Large-Scale Dataset for Training Next Generation Image-Text Models," in Advances in Neural Information Processing Systems (NeurIPS), vol. 35, pp. 25278–25294, 2022.
[7] A. Radford, J. W. Kim, C. Hallacy, et al., "Learning Transferable Visual Models From Natural Language Supervision," in Proc. Int. Conf. on Machine Learning (ICML), PMLR, vol. 139, pp. 8748–8763, 2021.
[8] O. Ronneberger, P. Fischer, and T. Brox, "U-Net: Convolutional Networks for Biomedical Image Segmentation," in Proc. Medical Image Computing and Computer-Assisted Intervention (MICCAI), pp. 234–241, 2015.