Controllable Zero-Shot Photo Customization through Scene-Aware Inversion

Abstract

Controllable Image Editing is advancing through efforts to align performance with evidence quality, resource limits, and transfer across settings. The discussion is organized around preserving identity and scene structure while exposing useful controls without per-image training. It synthesizes 1 focal paper with 11 independently retrieved publications reviewed against traceable publication metadata. The analysis is organized around inversion fidelity, edit locality, scene consistency, control strength, and evaluation. To avoid reading results from heterogeneous studies as exchangeable, the review compares problem boundaries, design logic, and conditions of validation. Across the literature, the comparison suggests that advances in controllable image editing become credible when representation, objective, and evaluation protocol are evaluated together and when uncertainty about distribution shift is reported explicitly. The resulting account aligns method selection to downstream risk and the conditions that weaken external validity, and proposes a research agenda centered on clear comparators, adversarial conditions, and inspectable evidence.

Keywords

Controllable Image Editing; Inversion Fidelity; Edit Locality; Scene Consistency; Control Strength; Evaluation

1. Introduction

Scholarship concerning controllable image editing must negotiate scientific ambition and practical constraint. A mature account offers not only stronger nominal performance but also explanations that locate performance within its operating envelope. That challenge is especially visible in preserving identity and scene structure while exposing useful controls without per-image training. An advantage observed in one experimental medium, dataset, cohort, geography, or system load can diminish when the evaluation environment moves. Evidence integration should therefore preserve the unit of analysis and the decision context. The review additionally distinguishes a mechanism from a correlation, a benchmark advantage from a deployment advantage, and an interpretable diagnostic from a causal explanation.

Five lenses organize the comparison: inversion fidelity, edit locality, scene consistency, control strength, evaluation. This set is useful because they jointly cover the technical object, measurement chain, and prerequisites of external validity. The synthesis is structured by representation, objective, and evaluation protocol. Under that principle, methodological novelty is necessary but insufficient; the comparison also evaluates comparator fairness, uncertainty reporting, and real operating constraints. The work reported here is a critical review and contains no newly asserted sample, experiment, measurement, or benchmark outcome.

2. Methodological Foundations

Four linked elements clarify the problem: the input evidence, the transformation applied to it, the output used for evaluation, and the decision that follows. In controllable image editing, each element is often assessed independently even though errors propagate across them. Bias in measurement can be carried forward and enlarged by the analytical method; an apparently accurate output can still be poorly calibrated for the final decision. Accordingly, ablation evidence should accompany aggregate performance. Sound comparative work specifies what the protocol fixes and what it lets move, then selects metrics that correspond to the intended use rather than to convenience alone.

Scope discipline is equally important. Inversion fidelity defines the local design problem, while evaluation determines whether the design can survive outside that boundary. Intermediate lenses such as edit locality and scene consistency connect those ends by exposing trade-offs that a single score conceals. Here, adverse outcomes and robustness analysis identifies the envelope that supports the interpretation. For applied work, documentation of deployment constraints is therefore part of the scientific result, not an administrative afterthought.

3. Comparative Evaluation

The target study U0612-01, CusEnhancer: A Zero-Shot Scene and Controllability Enhancement Method for Photo Customization via ResInversion [1], addresses a concrete part of controllable image editing through zero-shot residual inversion for scene-preserving, controllable photo customization. It identifies a representation, mechanism, or evaluation boundary needed to compare claims. Against the focus on preserving identity and scene structure while exposing useful controls without per-image training, the paper is interpreted within its documented design envelope: transfer may depend on its sensing regime, preparation, workload, population, or benchmark. The synthesis records the design boundary before comparing assumptions across sources.

Across the focal studies, inversion fidelity should be interpreted jointly with edit locality. A design can improve one while imposing a less visible trade-off on the other. The evidence can be organized in a matrix whose rows are operating conditions and whose columns are evidence quality, computation or process cost, robustness, and consequence of failure. This representation helps prevent an attractive mechanism from being detached from the circumstances that support it. It further reveals which ablations are informative: each ablation should answer why a component matters.

Scene consistency translates method behavior into decision evidence. A minimally complete evaluation should disaggregate routine and adverse conditions while making variability visible, and test plausible alternative explanations. Where distribution shift is central, calibration or sensitivity is as important as ranking. Where costs or hazards are asymmetric, utility-weighted assessment is more revealing than undifferentiated accuracy. These choices do not make studies identical; they make the reasons for their differences auditable.

4. Architecture and Learning Objectives

The design space is broadened by Diff-PC: Identity-preserving and 3D-aware controllable diffusion for zero-shot portrait customization [2]; ZeroScene: A Zero-Shot Framework for 3D Scene Generation from a Single Image and Controllable Texture Ed... [3]; Plot’n Polish: Zero-Shot Story Visualization and Disentangled Editing with Text-to-Image Diffusion Models [4]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of controllable image editing. Together they illuminate inversion fidelity: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

A first comparison set connects CDZL: a controllable diversity zero-shot image caption model using large language models [5]; Self-Attention Diffusion Models for Zero-Shot Biomedical Image Segmentation: Unlocking New Frontiers in... [6]; Attribute binding enhancement improvement for diffusion model based zero-shot image segmentation [7]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of controllable image editing. Together they illuminate edit locality: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

The neighboring evidence base brings together A Unified Approach for Conventional Zero-Shot, Generalized Zero-Shot, and Few-Shot Learning [8]; Enhancing Representation Inversion and Alignment for Zero-Shot Composed Image Retrieval [9]; Harmonizing Visual and Textual Embeddings for Zero-Shot Text-to-Image Customization [10]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of controllable image editing. Together they illuminate scene consistency: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

A complementary strand is visible in Zero-Shot Low-Light Image Enhancement via Diffusion with Joint Frequency and Spatial Guidance [11]; Integrating Latent and Pixel Diffusion Models for Zero-Shot Image Inpainting [12]. These publications were retrieved and verified through DOI or publisher metadata, and their titles directly intersect the problem boundary of controllable image editing. Together they illuminate control strength: some emphasize representations or mechanisms, whereas others foreground validation and application conditions. The useful comparison is not a leaderboard across incompatible settings, but a structured reading of what each study controls, leaves variable, and treats as a meaningful outcome. Divergence can then reveal a change in scale, data process, endpoint, or operating constraint.

5. Deployment and Governance

A practical workflow has five stages. First, define the decision and its failure cost. Second, specify the evidence and operating envelope. Third, choose a method whose inductive assumptions match that evidence. Fourth, test control strength and evaluation under controlled perturbations. Finally, retain provenance so that an unexpected outcome can be traced to data, configuration, material batch, environmental condition, or user interaction. This ordering holds preserving identity and scene structure while exposing useful controls without per-image training connected to observable requirements.

The systems-level lesson is that responsible progress in controllable image editing is governed by interfaces as well as components. Interfaces determine how uncertainty moves between measurement and inference, between algorithm and operator, or between microscopic design and system behavior. Teams should establish beforehand both success criteria and evidence for correction. When those agreements are explicit, efficiency goals remain subordinate to explicit validity, safety, and interpretation constraints.

6. Limitations and Future Directions

Interpretation is bounded by cross-study variation in labels, design choices, and reporting precision. Verified authorship and venue do not substitute for independent empirical replication. Fast-moving records create a further version-control problem. Scholar-confirmed focal citations were protected and metadata conflicts were surfaced rather than hidden.

Future work should preregister comparison protocols where feasible, make configurations computable and evaluate one meaningful change in context in deployment constraints. Research should also group adverse cases by process and consequence. For controllable image editing, a valuable shared benchmark would combine routine performance, deployment demand, calibration, and robustness after disruption. Such a benchmark would reward designs that remain useful when evidence is incomplete or conditions change.

7. Conclusion

Scholarship in controllable image editing is most informative when its studies are read relationally rather than a collection of isolated scores. The focal studies show how preserving identity and scene structure while exposing useful controls without per-image training; the additional literature reveals the assumptions required to interpret those contributions. Across inversion fidelity, edit locality, scene consistency, control strength, and evaluation, alignment remains the central requirement: the representation or mechanism must match the problem, metrics should fit the choice and real-world claims the evidence boundary. The consequence is evidence that travels more safely and fails more informatively.

References

1. Ren, M., Vaddamanu, P., Xu, J., & Frade, F. D. L. T. (2025). CusEnhancer: A Zero-Shot Scene and Controllability Enhancement Method for Photo Customization via ResInversion. arXiv preprint arXiv:2509.20775.

2. Xu, Y., Zhai, B., Zhang, C., Li, M., Li, Y., & Du, S. (2025). Diff-PC: Identity-preserving and 3D-aware controllable diffusion for zero-shot portrait customization. Information Fusion, 117, 102869. https://doi.org/10.1016/j.inffus.2024.102869

3. Tang, X., Li, R., & Fan, X. (2026). ZeroScene: A Zero-Shot Framework for 3D Scene Generation from a Single Image and Controllable Texture Editing. Computer Graphics Forum. https://doi.org/10.1111/cgf.70419

4. Akdemir, K., Shi, J., Kafle, K., Price, B. L., & Yanardag, P. (2026). Plot'n Polish: Zero-Shot Story Visualization and Disentangled Editing with Text-to-Image Diffusion Models. Proceedings of the AAAI Conference on Artificial Intelligence, 40(3), 1694-1702. https://doi.org/10.1609/aaai.v40i3.37147

5. Zhao, X., Kong, W., Liu, Z., Wang, M., & Li, Y. (2025). CDZL: a controllable diversity zero-shot image caption model using large language models. Signal, Image and Video Processing, 19(4). https://doi.org/10.1007/s11760-025-03871-9

6. Hamrani, A., & Godavarty, A. (2025). Self-Attention Diffusion Models for Zero-Shot Biomedical Image Segmentation: Unlocking New Frontiers in Medical Imaging. Bioengineering, 12(10), 1036. https://doi.org/10.3390/bioengineering12101036

7. Wu, N. (2024). Attribute binding enhancement improvement for diffusion model based zero-shot image segmentation. Applied and Computational Engineering, 86(1), 116-124. https://doi.org/10.54254/2755-2721/86/20241585

8. Rahman, S., Khan, S., & Porikli, F. (2018). A Unified Approach for Conventional Zero-Shot, Generalized Zero-Shot, and Few-Shot Learning. IEEE Transactions on Image Processing, 27(11), 5652-5667. https://doi.org/10.1109/tip.2018.2861573

9. Zhang, F., Chen, J., Zhou, H., & Xu, C. (2026). Enhancing Representation Inversion and Alignment for Zero-Shot Composed Image Retrieval. IEEE Transactions on Multimedia, 1-13. https://doi.org/10.1109/tmm.2026.3702526

10. Song, Y., Kim, J., Park, W., Shin, W., Rhee, W., & Kwak, N. (2025). Harmonizing Visual and Textual Embeddings for Zero-Shot Text-to-Image Customization. Proceedings of the AAAI Conference on Artificial Intelligence, 39(19), 20549-20557. https://doi.org/10.1609/aaai.v39i19.34264

11. Chu, J., Yu, X., & Lu, W. (2026). Zero-Shot Low-Light Image Enhancement via Diffusion with Joint Frequency and Spatial Guidance. Applied Sciences, 16(16), 8329. https://doi.org/10.3390/app16168329

12. Kakinuma, Y., Miyata, T., Hosono, K., & Kinoshita, H. (2026). Integrating Latent and Pixel Diffusion Models for Zero-Shot Image Inpainting. Nonlinear Theory and Its Applications, IEICE. https://doi.org/10.1587/nolta.2026ncp0007