Beyond Appearance: Evidence-Guided Programmatic 3D Scene Reconstruction with Agentic Lighting Reasoning

Zeting Luo Yifan Li Shuai Yang
Wangxuan Institute of Computer Technology, Peking University
2400013131@stu.pku.edu.cn, liyifan02@stu.pku.edu.cn, williamyang@pku.edu.cn



(a) Precise image-to-code reconstruction **(b) Rich applications: object insertion, texture editing and relighting**

Figure 1. **Lighting-aware image-to-code reconstruction.** Our framework reconstructs executable 3D scenes with explicit geometry, materials, and lighting, supporting diverse scene edits with realistic lighting interactions.

Abstract

Image-to-code methods for 3D scene reconstruction produce executable code that generates editable digital content and supports visual simulation. However, existing methods primarily rely on appearance similarity, which provides limited guidance for distinguishing errors in geometry, materials, and illumination. Light shapes scene appearance through its interactions with geometry and materials, influencing shading, cast shadows, and reflections. We therefore present ***LightHarness****, a VLM-based agentic framework that incorporates explicit lighting reasoning into iterative scene reconstruction. To be specific, we leverage complementary geometric, intrinsic, and illumination cues to jointly refine scene structure and appearance, and propose Lighting Reasoning Memory to maintain light-source hypotheses, which supports evidence, and revision history throughout iterations. Together, they support reasoning about partially visible and unseen sources, with evidence-guided verification directing subsequent revisions. Our framework reconstructs explicit scenes through coherent reasoning about geometry, materials, and illumination without additional training or differentiable rendering. Extensive experiments demonstrate superior image-to-code reconstruction accuracy under challenging shadows and complex illumination, further supporting downstream scene editing, object insertion, and versatile relighting.*

1. Introduction

Reconstructing an executable, editable 3D scene from a single image connects visual understanding with digital scene creation. Image-to-code reconstruction offers a promising direction: a scene program (*e.g.*, Blender, Three.js) explicitly represents geometry, spatial layout, materials, cameras, and light sources, enabling novel-view rendering and subsequent scene editing [10, 29, 35]. Except for spatial grounding that concerns object poses or location, **lighting** is also central to this process. Shading, cast shadows, highlights and reflections arise from the complex interactions between illumination, geometry, and materials. Accurate image-to-code reconstruction requires recovering object structure and layout together with the illumination that explains their observed appearance.

Recent advances in vision-language models (VLMs) [9, 24] have enabled scene construction through executable programs, connecting visual and semantic reasoning with explicit control over scene components. This capabil-

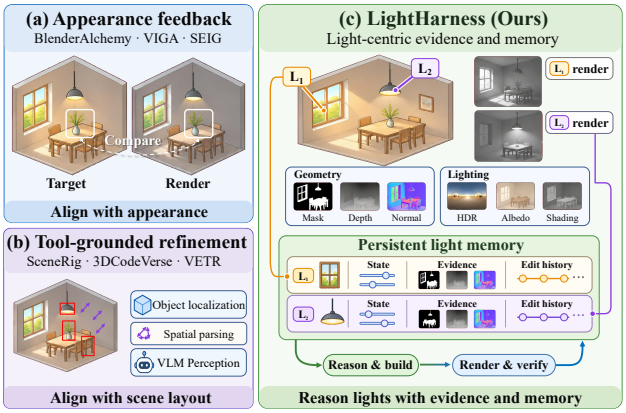


Figure 2. **Conceptual comparison of image-to-code paradigms.** **(a)** Appearance feedback emphasizes rendered–reference similarity. **(b)** Tool-grounded refinement uses perception and spatial analysis to improve layout. **(c)** LightHarness combines enriched evidence and persistent memory for explicit lighting reasoning throughout reconstruction.

ity supports a common iterative paradigm: an agent proposes or edits a scene program, inspects the rendered result, and revises its proposal based on visual feedback. SceneCraft [11] and BlenderAlchemy [12] demonstrate this paradigm to be effective in text-driven construction and appearance-guided editing, respectively. Extending it to reconstruction from visual input enjoys richer editing applications, and requires the VLMs to determine which scene configuration could explain the observed image. Recent systems formulate this problem through plan and observation management [35], staged refinement of scene attributes [10], or explicit geometric, material, and light modeling [4, 26] in synthetic game scenes.

Across these approaches, rendered-reference comparisons provide a central feedback signal for scene revision. However, such a discrepancy on appearance does not uniquely identify its underlying cause, because similar image differences can arise from geometry, reflectance, or illumination. For example, a dark region may arise from low surface reflectance, a surface facing away from light sources, or insufficient illumination from an unseen source. Simply brightening the material can reduce a local image discrepancy, but leave the underlying illumination error unresolved, leading to an incorrect material estimation. The challenge is therefore to connect observed lighting effects to their plausible geometric, material, and illumination explanations. This motivates our central question: **How can we explicitly reason about scene geometry, materials, and light sources for more accurate image-to-code reconstruction beyond appearance?**

To mitigate this, we propose **LightHarness**, a VLM-based agent framework that **incorporates explicit lighting modeling into iterative scene generation and verification with complementary evidence and reasoning**. We first associate light-source hypotheses with evidence of their interactions with scene geometry and materials. Specifically, we combine estimated HDR illumination [3], depth and surface orientation [13, 27], intrinsic albedo and shading [13]. These complementary sources of evidence relate appearance discrepancies to candidate revisions of the underlying scene factors, helping the *generator agent* determine which geometry, material, or light parameters to update during reconstruction, and finally output an executable Blender program. As for the *verifier agent*, we equip it with a rubric scorer and a VLM. The rubric scorer assesses rendered-reference agreement based on predefined criteria, while the VLM uses these assessments, reference-image history revisions, and decomposed rendering evidence to diagnose and propose refinement updates.

To preserve and refine illumination reasoning iteratively in the proposed agentic system, we design **Lighting Reasoning Memory** (LRM), shared by the generator and verifier agents. LRM associates each hypothesized light source with a group of stable parameters, including identity, visibility, location, color, *etc.*, establishing a clear correspondence among observations, editable Blender elements and the revision history. **Notably**, we explicitly consider partially visible, occluded, and off-frame light sources as hypotheses, enabling a more complete model of scene illumination that accounts for both visible and inferred light sources. In each iteration, the generator uses this persistent lighting proxy to build scene programs and refine parameters, while the verifier examines it based on rendered evidence and recommends corrections. By incorporating source-specific observations, decisions, and their supporting evidence in a synergistic framework, LRM provides persistent illumination-centric context for **LightHarness**, enabling subsequent geometry, material, and lighting revisions to build coherently on historical judgments.

Extensive experiments demonstrate our competitive 3D image-to-code reconstruction accuracy. We further verify our effectiveness on hard cases under complex lights and shadows, and showcase that our method can generalize to rich downstream applications, including image relighting, material/texture editing, and harmonious object insertion. Our contributions can be summarized as follows:

- We propose a VLM-based 3D image-to-code reconstruction framework that incorporates explicit lighting modeling to guide the recovery of geometry, materials, and light sources in an executable scene program.
- We introduce Lighting Reasoning Memory to preserve source identities, supporting evidence, and revision history across generation and verification, combining geometric and intrinsic priors with light-transport evidence to guide iterative scene refinement.
- We demonstrate competitive 3D image-to-code recon-

struction accuracy and further illustrate the utility of the reconstructed scenes through downstream applications, including image relighting, enhancement and complex light understanding.

2. Related Work

Lighting Estimation. Inverse rendering [14] aims to estimate scene geometry, reflectance, and illumination to explain observed image appearance and enable editing [15]. PhotoScene [33] optimizes materials and lighting on approximately aligned CAD geometry, while Materialist [28] combines intrinsic prediction with differentiable material and environment-light refinement. Multi-view methods further constrain material and illumination recovery through shared observations [18, 32]. LUMEN [3] instead predicts local HDR illumination, providing a lighting prior rather than a complete scene program. Generative relighting also incorporates physical guidance and local control through light-transport consistency [36], intrinsic conditioning and scribbles [6], or estimated illumination priors [34]. These methods provide complementary forms of lighting estimation and control, with different assumptions about available geometry and observations. Our focus is to use such evidence to guide geometry, material, and light revisions during executable scene reconstruction, with relighting as a downstream application.

Executable Scene Representations and Agentic Reconstruction. SceneCraft [11] translates scene descriptions and spatial constraints into Blender programs, while BlenderAlchemy [12] searches scene edits through rendered feedback. For image-to-code reconstruction, VIGA [35] combines generation and verification with multimodal memory, and SEIG [10] refines geometry, materials, composition, and lighting in stages. HARMONY [26] and Lumera [4] incorporate geometric priors and explicit illumination, with Lumera retaining shared refinement history. VETR [21] additionally uses execution observations to distinguish proposed edits from their realized effects. Notably, Code Video Model [17] combines executable codes with frontier generative models, achieving versatile controllable video generation. Although these systems already incorporate multiple forms of scene evidence and iterative feedback, our focus is to maintain source-specific lighting hypotheses and connect memorable evolution to geometric, intrinsic, and decomposed rendering evidence throughout reconstruction.

3. Method

3.1. Motivation and Overview

Image-to-code reconstruction represents geometry, materials, and lighting in an executable program, providing an interpretable proxy for rendering, scene manipulation, and downstream applications. Existing methods primarily exploit this problem through scene composition, spatial constraints, and interactive environments [8, 11, 19]. However, faithful reconstruction requires understanding how scene components interact with **light** to produce the observed appearance. Inferring lighting from incomplete observations is ill-posed because different scene configurations can produce similar images. A dark wall may indicate low reflectance, occlusion, or a surface facing away from illumination. Misinterpreting these effects can lead to incorrect materials or geometry, hampering reconstruction fidelity and subsequent editing. We therefore treat lighting reasoning as a crucial part of image-to-code reconstruction.

As shown in Fig. 3, **LightHarness** formulates a T-round loop which follows *Perceive* $\rightarrow$ *Generator* $\rightarrow$ *Execute & Render* $\rightarrow$ *Verifier* and combines visual evidence E extracted from image I with Lighting Reasoning Memory (LRM) M_t. Perception forms the scene context, which guides the Generator in producing code C_t and light-source hypotheses. Following rendering, rubric scoring and VLM diagnosis assess the reconstruction and guide subsequent revisions. Generated scene states and verification findings update LRM, preserving evidence and reasoning across rounds. The best valid candidate is selected by its rubric score, yielding jointly refined geometry, materials, and lighting for a wide range of downstream applications.

3.2. Enriched Evidence for Lighting Reasoning

Lighting reasoning requires both the spatial structure governing light transport and evidence of its effects on surfaces. We extract complementary geometric, intrinsic, and lighting cues to support these judgments.

Geometric evidence. SAM 3 [2] provides fine-grained semantic object masks, while VGGT [27] estimates relative depth and cameras to establish a common reconstruction coordinate system. For planar structures, we back-project reliable masked depth pixels and fit planes through robust estimation using least-squares refinement. Each plane provides a surface point and normal for reasoning about position, orientation, and illumination. Non-planar objects retain masks and depth cues without planar fitting.

Intrinsic and lighting evidence. Marigold [13] estimates albedo and diffuse shading to help distinguish reflectance from light effects. LUMEN [3] offers an HDR image suggesting light-source directions and chromatic properties. Evidence is extracted once and remains fixed, while its interpretation and reasoning evolve with rendering feedback.

3.3. Lighting Reasoning Memory

Scene revisions can change a light source's contribution even when its parameters remain unchanged. Lighting reasoning must therefore retain both current source states and the evidence underlying previous decisions. Therefore, we

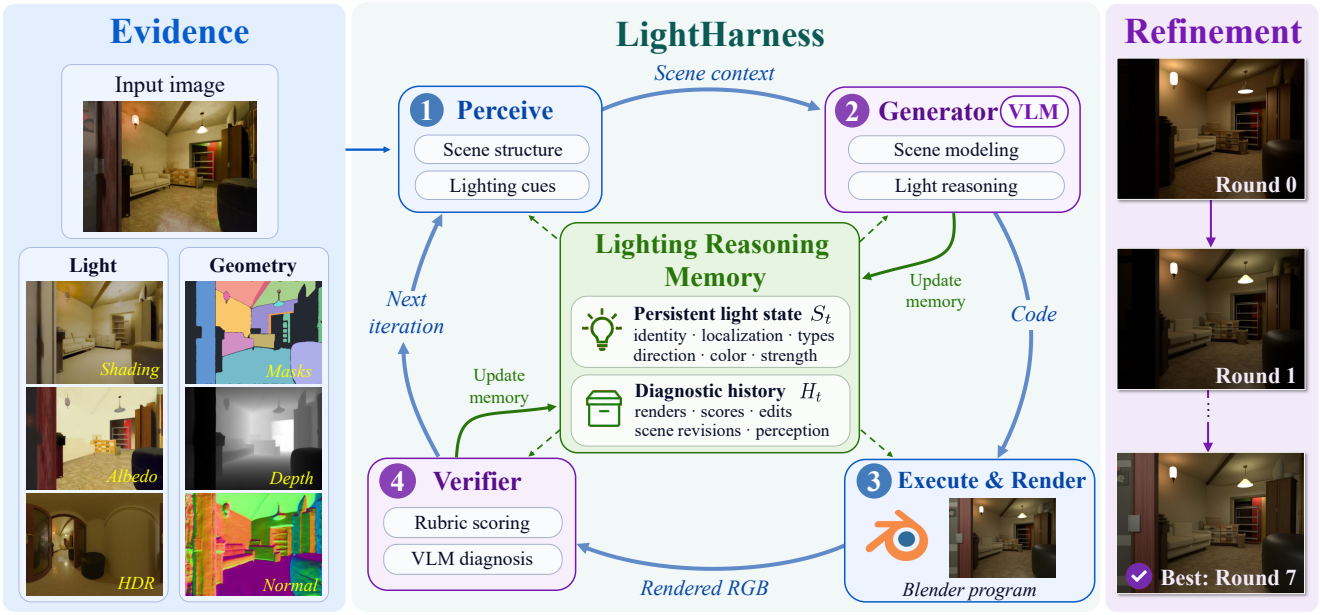


Figure 3. **Overall pipeline of LightHarness.** Geometric and lighting evidence guides a closed loop of perception, VLM-based generation, execution and rendering, and verification. Rubric scoring and VLM diagnosis guide joint scene refinement, while Lighting Reasoning Memory preserves source states and diagnostic history across iterations.

propose shared Lighting Reasoning Memory (LRM):

$$M_t = (S_t, H_t), \tag{1}$$

where S_t contains light-source states and H_t records diagnostic history. We connect the stable light-source identities with corresponding Blender elements, where each state describes light visibility, type, position, orientation, intensity and color. VLM visual reasoning [9, 24], guided by enriched evidence, also supports hypotheses of partially visible, occluded, and off-frame light sources.

The history H_t associates sources with previous assessments, proposed revisions, implemented changes, and observed outcomes. It further relates multiple sources through semantic correspondence to support a coherent interpretation of scene illumination. Together, S_t and H_t guide subsequent geometry, material, and lighting reasoning. Their update is described as below.

3.4. Evidence-Guided Verification

The Verifier consists a deterministic rubric scorer and a VLM diagnosis to assess reconstruction quality and guide scene refinement. The scorer measures agreement with the reference image and enriched evidence, while the VLM interprets these measurements alongside rendered outputs and LRM to identify plausible causes of discrepancies.

Rubric-based assessment. The rubric is six-fold as below. (1) *Object-mask error* measures projected-region disagreement. (2) *Relative-depth error* compares reliable overlapping regions after positive scale alignment and penalizes missing coverage. (3) *Orientation error* compares depth-derived plane normals while accounting for sign ambiguity. (4-5) *Albedo and diffuse-shading errors* are measured in trusted diffuse regions in linear color space. Shading uses one positive alignment scale shared across channels to preserve chromatic differences. (6) *RGB error* evaluates overall appearance. We aggregate these errors into scores q_t:

$$q_t = \frac{\sum_{k\in\mathcal{K}} w_k e_{t,k}}{\sum_{k\in\mathcal{K}} w_k}, \tag{2}$$

where $\mathcal{K}$ indexes mask, depth, orientation, albedo, shading, and RGB errors. Corresponding weights w_k remain fixed across rounds at $(2, 1, 0.5, 0.5, 1, 1)$, with sensitivity examined through ablations provided in the Supp. Lower scores indicate better agreement and guide candidate selection without gradient-based optimization.

Evidence-guided diagnosis. Beyond scalar assessment, the VLM combines component errors, rendered passes, source contributions, and LRM to recommend geometric, material, or lighting revisions. For an underlit wall, it examines orientation, occlusion, albedo, and relevant source contributions. Geometric discrepancies motivate structural revisions, whereas consistent geometry and albedo direct attention to lighting. Each recommendation specifies the discrepancy, proposed revision, and supporting evidence, providing guidance for the subsequent reconstruction round.

3.5. Memory-assisted Scene Reconstruction Loop

With enriched evidence, LRM, and the Verifier defined, we integrate them into a closed loop of perception, generation,

execution/rendering, and verification. At round t, the perception module combines the input image I, fixed evidence E, and memory M_t into a scene context comprising structural cues, lighting hypotheses, and unresolved discrepancies. Conditioned on this context and preceding scene code, the Generator jointly reasons about structure and illumination to produce t-step code C_t and a light source state update ΔS_t. After checking state consistency and source correspondence, we obtain

$$(C_t, \Delta S_t) = \text{VLM}(I, E, M_t, C_{t-1}), \quad \widetilde{S}_t = \text{Apply}(S_t, \Delta S_t). \tag{3}$$

The generated geometry and materials are combined with $\widetilde{S}_t$ under fixed camera and rendering settings. Blender $\mathcal{R}$ then renders the candidate scene as

$$Y_t = \mathcal{R}(C_t, \widetilde{S}_t), \tag{4}$$

including RGB, masks, depth, surface information, albedo, diffuse shading, and individual light sources and world-illumination contributions. Masks and depth show scene layout and occlusion. Albedo and shading help separate material color from lighting, while light source attribution indicates which region it affects. The Verifier assesses these outputs against the reference evidence, producing the rubric score q_t and evidence-guided recommendations.

Following each valid round, we reorganize the light source state and append a record h_t containing the diagnosis, implemented changes, scores, and references to archived outputs:

$$M_{t+1} = (\widetilde{S}_t,\ H_t \parallel h_t), \tag{5}$$

where $\parallel$ denotes history concatenation. The updated memory guides perception and generation in the next round, allowing subsequent reasoning to build on previous evidence and outcomes.

Since revisions may improve some factors while degrading others, we select

$$t^* = \arg\min_{t \in \mathcal{T}_{\text{valid}}} q_t \tag{6}$$

over completed valid rounds after T iterations. We return $(C_{t^*}, \widetilde{S}_{t^*})$ together, preserving the correspondence between the selected executable scene and its lighting state. This closed loop connects evidence, diagnosis, and persistent memory to support coherent refinement of geometry, materials, and illumination.

4. Experiments

Our experiments compare reconstruction quality with recently blooming image-to-code methods, examine the roles of visual evidence and Lighting Reasoning Memory, and assess generalizable downstream application and hard cases.

4.1. Implementation Details

Framework development. We load external modules (*i.e.*, SAM3, VGGT, Marigold, LUMEN) on two NVIDIA 4090-48G GPUs, and use GPT-6-Astra [24] as the Generator/Verifier VLM via API. All reconstructions are done with 10 rounds. We also examine our framework under different VLMs in the Supplementary materials.

Data. We evaluate on 32 images drawn from diverse benchmarks from robotics, relighting and image enhancement, comprising 10 from Hypersim [25], 4 from LuxRemix [16], 10 from BAID [20], 7 from MIT-Adobe FiveK [1], and 1 from RoboDojo [5] with complex lighting conditions, diverse geometry and materials. The whole testing set contains 17 real photographs and 15 synthetic images, covering 22 indoor and 10 outdoor scenes.

Metrics. We evaluate source-view consistency using PSNR [30], SSIM [31], LPIPS [37], and DreamSim [7]. Following SceneRig [19], we assess 3D reconstruction quality from the source view and nearby views at $\pm 15°$. By orbiting the camera around the world vertical axis through a pivot located four scene units along the original camera's viewing direction, we use GPT-6-Astra as a judge to compute the mean-centered Bradley-Terry scores among these rendered pairs, with higher scores indicating stronger relative preference.

Compared methods. We compare with recent SoTA image-to-code reconstruction methods, including VIGA [35], 3DCodeVerse [8], SceneRig [19]. All these methods use GPT-6-Astra as the backbone VLM with the same loop iterations and API. We additionally compare GPT-6-Sol with Codex [22] harness.

4.2. Image-to-Code Reconstruction

Qualitative results. Fig. 4 compares reconstructions across diverse geometry and lighting conditions. Our method better captures window–foreground contrast, localized illumination around bedside lamps, and combined window and ceiling lighting in the hall. It also preserves furniture arrangement and foreground animal composition, while competing methods exhibit layout distortions, simplified geometry, or mismatched illumination, supporting joint reasoning about scene structure, materials, and lighting. **Notably**, our method infers unseen light sources and structures through their illumination effects. In the fifth row, it reproduces striped shadows across the bedroom wall and bedding while preserving their spatial relationships, whereas competing methods miss these shadows or misinterpret the layout. The cast shadows in the last row further illustrate lighting reasoning **beyond appearance**, enabling more faithful image-to-code reconstruction. We present more visualized results under nearby views in the Supp.

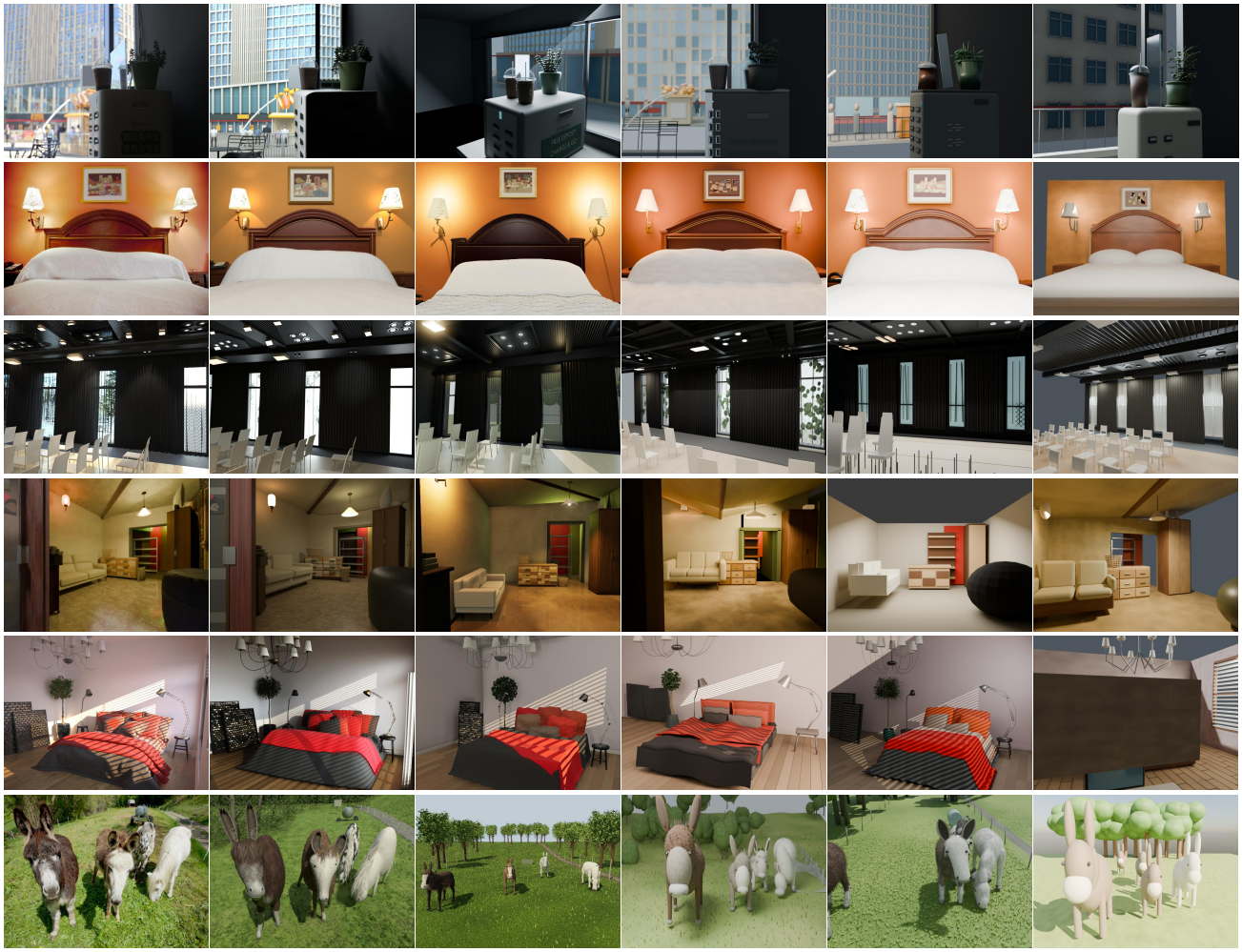


Figure 4. **Qualitative comparison between other SoTA image-to-code baselines.** Across diverse indoor and outdoor scenes, our **LightHarness** better preserves object arrangement and characteristic lighting effects, including strong back-lighting, localized illumination, and structured cast shadows, achieving more faithful reconstructions of scene structure and appearance.

Hard case verification. Fig. 5 compares reconstructions of a scene with a prominent diagonal cast shadow. While competing methods distort or omit this shadow, **LightHarness** better reproduces its spatial extent while preserving the surrounding building structure. This result supports the value of enriched evidence and persistent lighting memory in relating shadow patterns to scene geometry and illumination, enabling more faithful image-to-code reconstruction under challenging lighting.

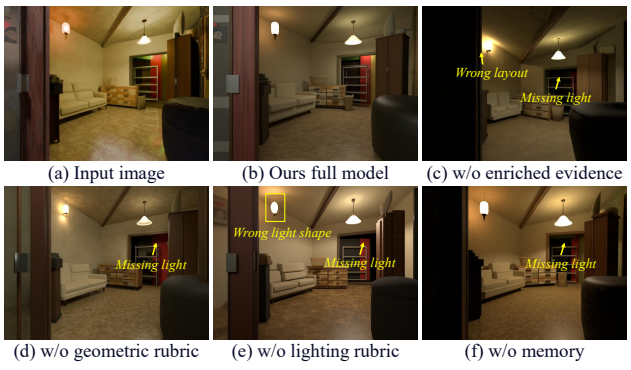
Quantitative results. Table 1 shows that our method achieves the best PSNR, LPIPS, and DreamSim among the compared methods. Relative to the recent SoTA method, VIGA, **LightHarness** improves PSNR by 0.75 dB and reduces DreamSim by 10.7%, indicating our superiority in both pixel-level fidelity and perceptual similarity. The VLM-based evaluation and user preference further demonstrate our advantages at the scene level, ranking it first in all criteria covering completeness, layout, and appearance. These consistent preferences support the effectiveness of our framework in reconstructing scene composition and visual appearance.

Inference cost. The latency mainly includes evidence extraction, GPT API and Blender rendering in the generation-verification loop. For a 768×576 image, these three parts cost 4.5 minutes, 118.8 minutes, and 1.1 minutes respectively. The latency is comparable with other baselines (VIGA [35] for 71.2 minutes and 3DCodeVerse [8] for 126.3 minutes). The critical bottleneck is API invocation, while other parts are relatively efficient with less than 6 minutes on two 4090-48 GB GPUs.

User Study. We invite 29 participants to rank the reconstructed scenes according to **(1)** scene structure fidelity and **(2)** lighting appearance fidelity, using the reference images as guidance. We achieve the first-place preference rate on



Figure 5. **Hard case verification.** Our method more faithfully reconstructs the diagonal shadow and surrounding building structure relative to other methods.



(a) Input image (b) Ours full model (c) w/o enriched evidence

(d) w/o geometric rubric (e) w/o lighting rubric (f) w/o memory

Figure 6. **Qualitative ablation of key designs.** Our full framework reproduces scene layout and illumination more precisely, while removing individual components introduces errors in basic scene layout (c), light placement (c-f), and light source shape (e).

both criteria, indicating stronger human preference for its reconstruction of scene structure and illumination.

4.3. Ablation Studies

We conduct quantitative and qualitative ablation experiments to demonstrate the effectiveness of each design in our framework in Fig. 6 and Table 2.

w/o enriched evidence. We disable the light and geometric evidence in Sec. 3.2, observing significant reconstruction error. As shown in Fig. 6 (c), the agent misjudges the wall light's distance from the surface and misses background illumination.



Figure 7. **Application of texture editing.** Our method supports texture replacement with multiple reference images with precise and executable code reconstruction, while directly applying even front-tier generative models leads to inaccurate structure.

w/o geometric rubric. Discarding explicit constraints on geometry features leads to incomplete scene light reconstruction, as shown in Fig. 6 (d).

w/o lighting rubric. Discarding lighting measurement produces wrong light source shapes and mismatched local illumination strength as shown in Fig. 6 (e).

w/o lighting reasoning memory. We compare our full model with a variant that retains only current scene code and parameter values. Without memory, background illumination is missed as shown in Fig. 6 (f), highlighting the value of explicitly organizing light source hypotheses and diagnostic history across iterations.

4.4. Generalizable Applications

Building upon the precise executable scenes we reconstruct, our framework supports diverse downstream applications, including scene texture editing, harmonious object insertion, versatile relighting, and degraded illumination restoration. We also demonstrate that our rendered edits can combine with the frontier generative model, GPT-Image [23] to produce more physically realistic results. Note that simply applying generative models is hard to achieve precise physical modeling like lighting interaction with complex shadows and textures. Our integration with GPT-Image enables advanced editing ability.

Material and Texture Editing. Our method can perform exact texture editing with multiple references in Fig. 7.

Versatile relighting with flexible control. Our method achieves superior light control relative to simply applying GPT-Image, as shown in Fig. 8 and Fig. 9.

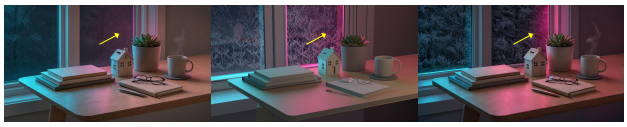
Harmonious object insertion. Our framework enables physically realistic object insertion as shown in Fig. 10.

Table 1. **Quantitative comparison of image-to-code reconstruction.** We report four appearance consistency metrics covering PSNR, SSIM, LPIPS and DreamSim, and measure novel-view consistency by a mean-centered Bradley-Terry (BT) score which is estimated from pairwise judgments using an additional $\pm 15^\circ$ view following SceneRig [19]. Bold and underline indicate the best and second-best.

	Source-view reconstruction				VLM evaluation (BT ↑)			User Preference ↑	
Method	PSNR ↑	SSIM ↑	LPIPS ↓	DreamSim ↓	Completeness	Layout	Appearance	Light	Structure
3DCodeVerse [8]	10.21	0.428	0.677	0.363	−0.367	−0.830	−0.594	6.5%	3.8%
GPT-6 Sol	13.03	<u>0.503</u>	0.548	0.318	−0.325	−0.284	−0.262	6.9%	2.5%
VIGA [35]	<u>14.28</u>	**0.525**	<u>0.446</u>	<u>0.242</u>	<u>0.524</u>	<u>0.537</u>	<u>0.700</u>	<u>21.4%</u>	<u>17.8%</u>
SceneRig [19]	11.86	0.455	0.665	0.420	−1.092	−1.043	−1.232	3.3%	1.3%
LightHarness (Ours)	**15.03**	0.486	**0.437**	**0.216**	**1.259**	**1.620**	**1.388**	**61.9%**	**74.6%**

Table 2. **Quantitative ablation experiment.** We evaluate on a 16-image subset for efficiency. Our full model achieves the best. BT scores are re-normalized here.

Variant	PSNR↑	DreamSim↓	BT↑ Compl.	BT↑ Appear.
w/o enriched evidence	10.05	0.351	−0.4864	−1.1035
w/o geometric rubric	13.56	0.278	−0.7148	−0.2589
w/o lighting rubric	13.70	0.239	0.3230	0.5107
w/o LRM	12.81	0.270	−0.0078	−0.0724
Full model	**14.24**	**0.213**	**0.8861**	**0.9241**



(a) Directly edit by GPT-Image (b) Edit on our recon. scene (c) Our editing with GPT-Img. post-process

Figure 8. **Application of relighting with realistic reflection.**



(a) Input image (b) Edit on our recon. scene

(c) Directly edit by GPT-Image (d) Our editing with GPT-Image post-process

Figure 9. **Application of relighting with realistic shadow.**

5. Conclusion, Limitation and Future Work

We propose LightHarness, a VLM-based agentic framework that integrates explicit lighting reasoning and persistent light-source memory into image-to-code 3D reconstruction. Complementary geometric, intrinsic, and illumination evidence guides reconstruction, while rubric-based assessment and VLM diagnosis support iterative refinement. Lighting reasoning memory preserves source identities, evidence, semantic relationships, and revision history, enabling coherent reasoning across rounds. Quantitative and qualitative results demonstrate our improved reconstruction fidelity under challenging shadows and complex illumination compared to recent SoTA methods. Based on the accurate programmatic scene reconstruction, our framework enables rich applications, including relighting, harmonious object insertion, material and texture editing with plausible physical modeling.



(a) Input image and reference obj. (b) Edit in our recon. scene

(c) Directly edit by GPT-Image (d) Ours render + GPT-Image editing

Figure 10. **Application of harmonious object insertion.** Our method **(b, d)** enables physically realistic lighting, reflection, occlusion and geometric projection, while **(c)** generative model-only method struggles with these real-world principles.

Limitation. Our framework remains limited in inferring unseen scene objects from sparse and overlapping shadow cues. As shown in Fig. 11, it fails to reconstruct the off-frame people and sign suggested by their shadows, leaving the corresponding shadow patterns incomplete.

Future work. We will incorporate complex scene examples through in-context learning and post-train open-source VLMs on challenging datasets to strengthen reason-



(a) Input image (b) Our rendered image

Figure 11. **Limitation on complex invisible lighting reasoning.** Although effective, our framework still struggles with inferring complex invisible objects from scarce light and shadow cues.

ing about unseen objects and their interactions with light. Reducing token budget is also meaningful. We argue that projecting visual representation back to executable codes is a highly promising direction in this field.

LLM Usage Statement

OpenAI Codex was used to assist with basic code generation, manuscript organization and language editing. Codex was not used to formulate research ideas, design the method or experiments, or draw scientific conclusions. All AI-assisted text was manually reviewed and revised by the authors. All AI-assisted code was inspected, executed, and verified for correctness, and the resulting visualizations and preprocessed data were checked against the source data and expected outputs. The authors take full responsibility for the final content, including all text, code, results, and claims.

References

[1] Vladimir Bychkovsky, Sylvain Paris, Eric Chan, and Frédo Durand. Learning photographic global tonal adjustment with a database of input / output image pairs. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2011. 5

[2] Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. In *Int. Conf. Learn. Represent.*, 2026. 3

[3] Hongze Chen, Fangyu Lin, Runze Zhang, Lingting Zhu, Weikai Chen, Zeyu Hu, Yingda Yin, Keyang Luo, Xin Wang, Zehong Lin, and Jun Zhang. LUMEN: Geometry-guided diffusion model for spatially-varying HDR illumination estimation. In *Siggraph Asia*, 2026. 2, 3

[4] Junhao Chen, Xinghao Chen, Henghaofan Zhang, Zihao Qiao, Saining Zhang, Yongzhi Li, Ruqi Huang, Sisi Li, Yimin Sheng, Jianyi Zhu, and Hao Zhao. Engine-native editable 3D world reconstruction with objects and lighting. *arXiv:2607.20889*, 2026. 2, 3

[5] Tianxing Chen, Yue Chen, Zixuan Li, Junyuan Tang, Kailun Su, Weijie Wan, Baijun Chen, Haoran Lu, Haowen Yan, Honghao Su, et al. RoboDojo: A unified sim-and-real benchmark for comprehensive evaluation of generalist robot manipulation policies. *arXiv:2607.04434*, 2026. 5

[6] Jun Myeong Choi, Annie Wang, Pieter Peers, Anand Bhattad, and Roni Sengupta. ScribbleLight: Single image indoor relighting with scribbles. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025. 3

[7] Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Advances in Neural Information Processing Systems*, 2023. 5

[8] Yipeng Gao, Ziyao Zeng, Jinfa Huang, Yijiang Li, Haolin Xiong, Wang Qin, Xinlei Yu, Youheng Yao, Jie Zhang, Kun-Yu Lin, Rong Liu, Yinyuan Zhao, Yajie Zhao, Meiqi Guo, Jiebo Luo, Lei Shu, Yunhao Ge, Xinyu Hu, and Laurent Itti. 3DCodeVerse: Exploring 3D worlds as code. *arXiv:2610.22341*, 2026. 3, 5, 6, 8

[9] Google DeepMind. Gemini 4 Carbon model card. Technical report, Google DeepMind, 2026. 1, 4

[10] Guangzhao He, Rundong Luo, Wei-Chiu Ma, and Hadar Averbuch-Elor. Thinking in blender: Staged executable inverse graphics with vision-language models. *arXiv:2606.02580*, 2026. 1, 2, 3

[11] Ziniu Hu, Ahmet Iscen, Aashi Jain, Thomas Kipf, Yisong Yue, David A Ross, Cordelia Schmid, and Alireza Fathi. SceneCraft: An LLM agent for synthesizing 3D scenes as blender code. In *Int. Conf. Mach. Learn.*, 2024. 2, 3

[12] Ian Huang, Guandao Yang, and Leonidas Guibas. BlenderAlchemy: Editing 3D graphics with vision-language models. In *Eur. Conf. Comput. Vis.*, 2024. 2, 3

[13] Bingxin Ke, Kevin Qu, Tianfu Wang, Nando Metzger, Shengyu Huang, Bo Li, Anton Obukhov, and Konrad Schindler. Marigold: Affordable adaptation of diffusion-based image generators for image analysis. *arXiv:2505.09358*, 2025. 2, 3

[14] Zhengqin Li, Mohammad Shafiei, Ravi Ramamoorthi, Kalyan Sunkavalli, and Manmohan Chandraker. Inverse rendering for complex indoor scenes: Shape, spatially-varying lighting and SVBRDF from a single image. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, pages 2475–2484, 2020. 3

[15] Zhengqin Li, Jia Shi, Sai Bi, Rui Zhu, Kalyan Sunkavalli, Miloš Hašan, Zexiang Xu, Ravi Ramamoorthi, and Manmohan Chandraker. Physically-based editing of indoor scene lighting from a single image. In *Eur. Conf. Comput. Vis.*, 2022. 3

[16] Ruofan Liang, Norman Müller, Ethan Weber, Duncan Zauss, Nandita Vijaykumar, Peter Kontschieder, and Christian Richardt. LuxRemix: Lighting decomposition and remixing for indoor scenes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2026. 5

[17] Zhiyang Liang, Jiawei Zhang, Guocheng Qian, Dongdong Chen, and Ziyu Wan. Code video model, 2026. Technical report. 3

[18] Chih-Hao Lin, Jia-Bin Huang, Zhengqin Li, Zhao Dong, Christian Richardt, Tuotuo Li, Michael Zollhöfer, Johannes Kopf, Shenlong Wang, and Changil Kim. IRIS: Inverse rendering of indoor scenes from low dynamic range images. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025. 3

[19] Rundong Luo, Yunong Liu, Di Cao, Matthew Tancik, Shyamal Buch, Colton Stearns, and Wenqi Xian. SceneRig: An agentic system for simulation-ready 3D scene reconstruction from single images. *arXiv:2610.28585*, 2026. 3, 5, 8

[20] Xiaoqian Lv, Shengping Zhang, Qinglin Liu, Haozhe Xie, Bineng Zhong, and Huiyu Zhou. BacklitNet: A dataset and network for backlit image enhancement. *Comput. Vis. and Image Understanding*, 218:103403, 2022. 5

[21] Yingbo Ma, Kuangshi Ai, Gordon Dai, Chenglong Yu, Chaoli Wang, and Hua Shen. VETR: Execution-grounded self-improving agent for symbolic graphics reasoning. *arXiv:2610.13098*, 2026. 3

[22] OpenAI. Codex. OpenAI Developers. 5

[23] OpenAI. Introducing our latest image generation model in the API. OpenAI Blog, 2025. 7

[24] OpenAI. GPT-6 Astra system card. Technical report, OpenAI, 2026. 1, 4, 5

[25] Mike Roberts, Jason Ramapuram, Anurag Ranjan, Atulit Kumar, Miguel Angel Bautista, Nathan Paczan, Russ Webb, and Joshua M. Susskind. HyperSim: A photorealistic synthetic dataset for holistic indoor scene understanding. In *Int. Conf. Comput. Vis.*, 2021. 5

[26] Shufan Sun, Chen Wang, Enxin Song, Jiatao Gu, and Lingjie Liu. HARMONY: Hierarchical agentic reasoning for monocular image-to-scene synthesis. *arXiv:2609.26793*, 2026. 2, 3

[27] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. VGGT: Visual geometry grounded transformer. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2025. 2, 3

[28] Lezhong Wang, Duc Minh Tran, Ruiqi Cui, Thomson TG, Anders Bjorholm Dahl, Siavash Arjomand Bigdeli, Jeppe Revall Frisvad, and Manmohan Chandraker. Materialist: Physically based editing using single-image inverse rendering. *Int. J. Comput. Vis.*, 134(6), 2026. 3

[29] Yuhao Wang, Puyi Wang, Linjie Li, Zhengyuan Yang, Kevin Qinghong Lin, and Yu Cheng. 3D-CoS: A new 3D reconstruction paradigm based on VLM code synthesis. *arXiv:2606.10478*, 2026. 1

[30] Zhou Wang and Alan C. Bovik. Mean squared error: Love it or leave it? a new look at signal fidelity measures. *IEEE Signal Processing Magazine*, 26(1):98–117, 2009. 5

[31] Zhou Wang, Alan C. Bovik, Hamid R. Sheikh, and Eero P. Simoncelli. Image quality assessment: From error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, 2004. 5

[32] Liwen Wu, Rui Zhu, Mustafa B. Yaldiz, Yinhao Zhu, Hong Cai, Janarbek Matai, Fatih Porikli, Tzu-Mao Li, Manmohan Chandraker, and Ravi Ramamoorthi. Factorized inverse path tracing for efficient and accurate material-lighting estimation. In *Int. Conf. Comput. Vis.*, 2023. 3

[33] Yu-Ying Yeh, Zhengqin Li, Yannick Hold-Geoffroy, Rui Zhu, Zexiang Xu, Miloš Hašan, Kalyan Sunkavalli, and Manmohan Chandraker. PhotoScene: Photorealistic material and lighting transfer for indoor scenes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2022. 3

[34] Chi-En Yen, Duy-Khanh Ngo, Wen-Wei Tang, Huu-Phu Do, Wen-Hsiao Peng, and Ching-Chun Huang. FreeLit: Paired-free indoor relighting via physics-guided diffusion. In *ACM Int. Conf. Multimedia*, 2026. 3

[35] Shaofeng Yin, Jiaxin Ge, Zora Zhiruo Wang, Chenyang Wang, Xiuyu Li, Michael J. Black, Trevor Darrell, Angjoo Kanazawa, and Haiwen Feng. Vision-as-inverse-graphics agent via interleaved multimodal reasoning. In *Eur. Conf. Comput. Vis.*, 2026. 1, 2, 3, 5, 6, 8

[36] Lvmin Zhang, Anyi Rao, and Maneesh Agrawala. Scaling in-the-wild training for diffusion-based illumination harmonization and editing by imposing consistent light transport. In *Int. Conf. Learn. Represent.*, 2025. 3

[37] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric. In *IEEE/CVF Conf. Comput. Vis. Pattern Recog.*, 2018. 5