EZCAM: DATA, INFRASTRUCTURE, AND POLICY FOR EMBODIED PHOTOGRAPHY

Anonymous authors
Paper under double-blind review



Figure 1: **EZCAM: data, infrastructure, and a reference camera policy.** (a) The collection contains 96,700 game and game-derived 3DGS video–pose trajectories, with 347,111 auxiliary Internet clips documented separately. (b) The infrastructure binds 49 assets to calibrated observations, complete-path execution checks, and traceable episode records. (c) A history-conditioned camera policy learns from Internet, game, and 3DGS data within a unified framework and executes one predicted increment before replanning. Scene imagery is illustrative.

ABSTRACT

Embodied photography asks an agent to move a camera, observe the resulting view, and retain a photograph under a limited interaction budget. Studying this capability requires both action-linked supervision and an environment that exposes the consequences of the agent's own decisions. We present EZCAM, connecting photographic trajectories, interactive camera infrastructure, and a learned reference policy. The corpus contains 96,700 video–pose trajectories from games and game-derived reconstructions, with 347,111 auxiliary Internet clips indexed separately. Policy training also incorporates Internet video. Replay audits distinguish improving endpoints from non-monotonic local motion, and a blinded study supports selected data-image pairs. The infrastructure provides 49 scene assets with explicit camera conventions, complete-path checks, and observation accounting. We evaluate seven trained configurations across all 49 scenes, using six starts per scene and 40 action requests per episode. The learned policy achieves the highest mean best-observed raw score gain of 0.5126, compared with 0.2817 for the strongest supplied visual-feature adapter. Its mean endpoint gain is negative, illustrating why retained-image quality and final-pose quality must be distinguished. These results establish a measurable data–interaction–policy pipeline and characterize the trained-model comparison. The evaluation reports training, development, and unseen scenes separately. Controlled ablations measure the effects of

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

branch supervision and recovery; independent policy-output preference remains unverified.

1 INTRODUCTION

A photographer improves a composition by moving before pressing the shutter. A step sideways changes occlusion, background, and perspective in ways that cropping an existing image cannot. For a camera agent, the next view is revealed only after an action is executed. We call this task *embodied photography*: acquiring a photograph through camera actions under an explicit observation budget.

Aesthetic judgment supplies a signal about a visible image. Learning photographic action additionally requires examples connecting observations, motion, and visual outcomes, together with a way to evaluate the consequences of the learner's own decisions. Aesthetic predictors (Murray et al., 2012; Kong et al., 2016; Talebi & Milanfar, 2018), camera-trajectory models (Zhang et al., 2025b; Li et al., 2024), and robotic photography (AlZayer et al., 2021) provide relevant components. Our question is how to connect supervision, executable interaction, and policy measurement in a common photographic task.

EZCAM organizes this connection through data, infrastructure, and a policy (Figure 1). Its 96,700 game and game-derived 3DGS trajectories pair videos with camera poses and scored planning views. The 347,111 auxiliary Internet clips provide additional estimated camera-motion supervision for BSCP; the resource size is distinct from the retained training subset. Collection joins increasing-score key views, but continuous replay need not improve at every frame. We therefore audit complete videos, local action windows, and selected-image human preferences separately. The interaction infrastructure binds 49 reconstructed assets to camera and allowed-motion conventions, checks complete translation segments, and records the observations actually produced by requested actions.

Our Branch-Supervised Camera Policy (BSCP) predicts camera-local actions from visual and executed-action history. It learns from Internet videos and game (3DGS) demonstrations within the unified EZCAM framework. Across the 49-scene collection, it attains the highest mean best-observed raw score gain among the seven trained configurations: 0.5126, compared with 0.2817 for the strongest supplied visual-feature adapter. All models share the same evaluation interface and protocol. We report training, development, and unseen-scene results separately to assess transfer beyond the adaptation scenes.

The measurements expose a distinction essential to the task: a controller can find a high-scoring image and subsequently leave that view. Our policy's mean endpoint gain is -0.3653, despite positive best-observed gain. This does not invalidate a retained photograph; it means that retaining an image, ending at a good pose, and returning for a new exposure are different outputs. Likewise, a gain in a proxy score is not independent evidence of human preference. We make these distinctions explicit rather than treating a single trajectory maximum as a complete account of photographic ability.

Contributions. (1) **Data:** a pose-associated photographic trajectory resource with full-video, short-window, and selected-image preference evidence. (2) **Infrastructure:** a budgeted camera interaction contract with complete-path execution checks and distinct retained-image and endpoint measurements. (3) **Policy:** a documented learned reference system and a seven-model evaluation over 294 starts comparing trained configurations. Matched-update supervision and same-checkpoint recovery ablations quantify their contributions under this protocol.

2 RELATED WORK

Aesthetics and composition. Image-rating datasets and quality models (Murray et al., 2012; Kong et al., 2016; Talebi & Milanfar, 2018; Radford et al., 2021) provide supervision for scoring photographs. Cropping and composition methods (Wei et al., 2018; Zeng et al., 2019; Yao et al., 2025) operate on captured content. EZCAM studies actions that change the available view. A predictor used to construct trajectories is not an independent validator of those trajectories, motivating a separate human comparison.

Camera control and embodied policies. Camera-trajectory generation and virtual cinematography (Zhang et al., 2025b; Li et al., 2024; Courant et al., 2024; Liu et al., 2024; Jiang et al., 2024) provide useful motion priors. Navigation and vision-language-action work (Anderson et al., 2018; Zhang et al., 2024; 2025a; Brohan et al., 2023; Kim et al., 2024; Black et al., 2024) studies action prediction from observations. Photography changes the objective from reaching a specified location to selecting a desirable view. Active vision and next-best-view planning (Bajcsy, 1988; Connolly, 1985; Zeng et al., 2020) also connect sensing and motion, usually through information acquisition. Robotic photography (AlZayer et al., 2021) directly motivates our application. We focus on the data, execution contract, and accounting needed to compare approaches under a shared camera budget.

View synthesis as an interaction environment. Neural rendering and 3D Gaussian Splatting (Mildenhall et al., 2020; Kerbl et al., 2023) make captured scenes queryable at new viewpoints. Capture datasets (Barron et al., 2022; Knapitsch et al., 2017; Hedman et al., 2018) and embodied simulators (Savva et al., 2019) offer complementary appearance and interaction resources. EZCAM distinguishes geometric permission, image coverage, and photographic quality; a valid pose guarantees neither of the latter.

3 The Embodied Photography Task

State, observation, and action. For scene e, let $x_t = (\mathbf{p}_t, R_t)$ be the camera-to-world pose, K_e the intrinsics, and $I_t = \mathcal{R}_e(x_t, K_e)$ the rendered image. The policy receives its declared visual and executed-action history and proposes a local increment $u_t = (\delta\mathbf{p}_t, \delta R_t)$. The geometric executor applies Eq. 5 or preserves the full pose after rejection. Uncharged future-view previews are disallowed; online score features and deterministic recovery must be declared separately.

Interaction budget. The reported PF41 evaluation uses one initial observation and 40 subsequent action requests, producing 41 observations. A stationary or rejected request still consumes a request and its resulting observation; 40 requests need not produce 40 pose changes. Recovery actions occupy the same budget, rather than adding steps. Model invocations and observation counts are separate costs. The earlier execution pilot uses seven requests and eight observations and is reported only as infrastructure validation.

Photographic outputs and metrics. Let $q_t = q(I_t) = s(I_t)$ be the image aesthetic score defined in Eq. 4, and let a declared selector σ return an already observed photograph. We distinguish

$$G_{\text{best}} = \max_{0 \le t \le T} q_t - q_0, \qquad G_{\text{end}} = q_T - q_0, \qquad G_{\text{sel}} = q_{\sigma(0:T)} - q_0. \tag{1}$$

The main table reports G_{best}, not a separately validated selector. Because the initial image is included, $G_{\text{best}} \ge 0$ by construction. Best-observed scoring measures the available visited set; it neither validates human preference nor executes a return to that pose. The reported AUC is

$$\text{AUC} = \frac{1}{T} \sum_{t=1}^{T} \left(\max_{0 \le j \le t} q_j - q_0 \right), \quad T = 40, \tag{2}$$

and $\text{SR}_{0.3}$ is the fraction with $G_{\text{best}} \ge 0.3$. These are raw score differences, without scene-percentile normalization.

Paired comparison and scope. All methods use six starts per scene over the same 49-scene collection. Let $\bar{G}_{m,e}$ average G_{best} across those starts. We report the scene-balanced mean and paired differences against a trained adapter b:

$$\Delta_{m,b} = \frac{1}{49} \sum_{e=1}^{49} (\bar{G}_{m,e} - \bar{G}_{b,e}). \tag{3}$$

With equal starts per scene this also equals the episode-mean difference. The aggregate includes training and development scenes and is not a held-out generalization estimate. Scene-level wins and losses complement the mean; we do not claim original-environment-clustered significance without the grouping information needed to estimate it. A new exposure after returning to a previous viewpoint would require an executed path and additional cost; physical reacquisition is outside this evaluation.

4 DATA

Aesthetic score. The aesthetic score averages LAION V1 (a_1), V2.5 ($a_{2.5}$), and ArtiMuse (a_{AM}) after nominal scale alignment:

$$s(I) = \frac{1}{3}\left[a_1(I) + \mathrm{clip}(a_{2.5}(I), 0, 10) + \frac{a_{\mathrm{AM}}(I)}{10}\right], \tag{4}$$

where each a_m averages that model's two scoring heads. The fixed mean serves view selection, trajectory annotation, replay analysis, and policy evaluation. Predictor supervision appears in Appendix A.1; the human study below tests agreement with viewer preferences.

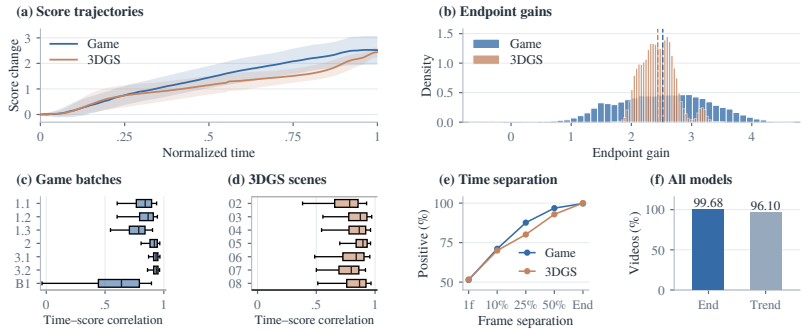

Figure 2: **Complete-corpus audit.** Score change, endpoint gain, temporal correlation, comparison rates, and component agreement. Definitions: Appendix B.

Human preference comparison. This study tests the scorer's alignment with human aesthetic preferences, independently of policy performance. All **37 participants** judged 60 blinded pairs (21 game, 39 3DGS; 2,220 responses), without agreement-based exclusion. Each pair compares the initial frame with the scorer's choice from 21 samples in the first 10 s. Higher-scoring images receive 1,417/1,609 directional choices (**88.07%**; two-way 95% bootstrap interval: 83.2–92.2%), or 63.83% of all 2,220 judgments; 462 are "similar" and 149 "cannot judge" (Figure 3). This supports an image-ranking proxy on the sampled pairs. Protocol and components: Appendices C and C.1.

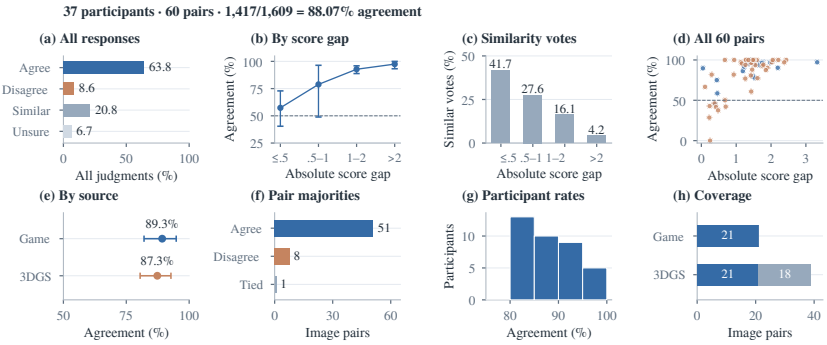

Figure 3: **Blinded comparison with 37 participants.** Responses and agreement; error bars: two-way 95% bootstrap intervals (Appendix C).

Sources and construction. The 96,700 trajectories pair video, camera poses, and scored planning views (Appendix A.2, Table 4). A custom UE5 runtime supports multi-game sampling and recording. A graph connects nearby, increasing-score views with bounded motion; route search varies approaches to each region's highest-scoring sampled view. The 96,700 deduplicated physical-pose sequences are distinct, although starts or goals can recur. Interpolation produces continuous replay; 3DGS views are rendered before the same search. BSCP also uses a retained SpatialVID subset (Wang et al., 2025) for visual and camera-motion supervision. Dataset and code release is described in Appendix A.5.

Replay and action windows. Across 44,264,749 scored frames, endpoint gains are positive in 99.74% of game and 100.00% of 3DGS videos (Figure 2). Interpolation changes compositions: about 51% of adjacent-frame comparisons are positive, versus 96.80% and 92.87% across half the frame-index span. At 2.13 s, exported-window improvement is 85.87% and 69.62%, respectively (Appendix B.4, Table 12). Thus better endpoints do not imply positive reward for every action. Appendix A.4 details training subsets and pose alignment.

Figure 4 shows endpoint images, camera paths, and scored planning views for four recorded examples.

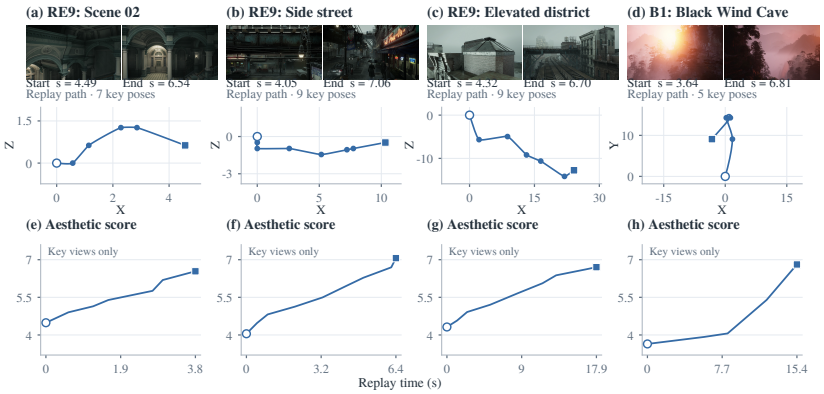

Figure 4: **Key views and replay paths.** (a–d) Endpoints and paths (circle: start; square: end). (e–h) Key-view scores, not intermediate-frame measurements. Video samples: Figure 12.

4.1 3DGS SCENE CONSTRUCTION

The 49 assets comprise 12 game, 15 interior, eight object-centric, four campus/aerial, and ten additional scenes. Captured scenes combine dense oblique images and viewpoint photographs for OpenFlyScan reconstruction (You et al., 2026), using recorded calibration for fixed-pose triangulation. Each bundle retains the Gaussian asset, calibration, coordinate transform, and allowed volume (Figure 5). A 128-pair KCD2 check yields 27.48 dB mean PSNR; views excluded from photometric fitting still participate in triangulation and quality checks. This assesses construction fidelity, not independent aesthetic ranking (Appendix A.6).

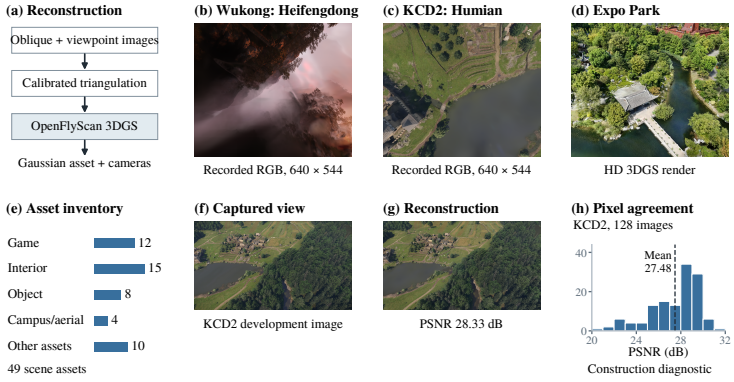

Figure 5: **Scene construction.** Reconstruction, example scenes, inventory, and captured/rendered comparisons. (h) 128 KCD2 development views; (d) Expo Park, cropped from a 1920×1080 render. Source pixels are unchanged.

5 INFRASTRUCTURE

The interface and verification evidence below describe the infrastructure and its development pilot. The completed 49-scene policy evaluation uses the final PF41 configuration in Appendix E; pilot-specific adapters and budgets are not the final policy protocol.

The environment connects assets, observations, allowed motion, and records for a point camera in a static scene (Figure 6). Geometry is available to the executor; policy access follows the adapter.

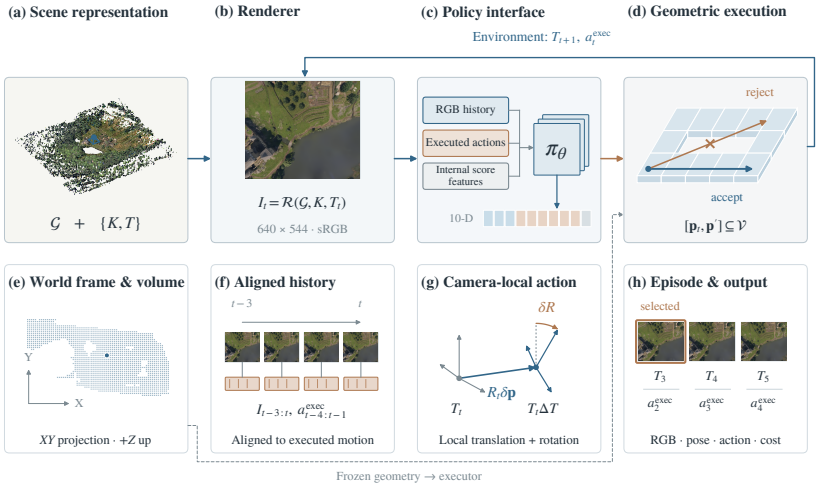

Figure 6: **Camera interaction infrastructure.** (a–d) Render–act loop; (e–h) coordinates, history, motion, and outputs. Imagery/geometry in (a,b,e,f,h) are recorded; policy and execution are schematic.

Scene and observation interface. World coordinates are right-handed, $+Z$ up; camera axes are right/down/forward. Alignment preserves native scale. EWA rendering returns 640×544, 90° horizontal-FOV, 8-bit sRGB with intrinsics, pose, and renderer identity. Logging does not grant policy access to state. Pilot permissions and preprocessing: Appendices D.3–D.4; policy configuration: Section 6.

Complete-path execution. Allowed cells define the runtime position domain. At pose $(\mathbf{p}, R)$, a local request gives

$$\mathbf{p}' = \mathbf{p} + R\delta\mathbf{p}, \qquad R' = R\delta R. \tag{5}$$

The translation must stay within allowed cells and avoid obstacles: sorted segment–cell intervals must cover $[0, 1]$. Rejection preserves pose. Execution uses neither scores nor candidate renders; a valid position may still face incomplete reconstruction. Geometry and native-unit decoding: Appendices D.2 and D.3.

Accounting and verification. Records separate requested/executed motion, pose, RGB hash, interventions, termination, and render/inference/query costs. Every reset or post-request observation counts, including stationary/rejected motion (Appendix D.3). Checks pass for 49 assets and 32 reference cases. The three-scene pilot has 1,008 transitions and 1,152 observations: all poses match independent recomputation; rerendering reproduces 576/576 images. Full-path checking rejects 108 of 192 longer segments with valid endpoints. Coverage and timing: Appendices D.5–D.6.

6 REFERENCE POLICY

History-conditioned camera actions. BSCP conditions on four RGB observations and four executed camera actions. An OpenCLIP ViT-L/14 encoder provides a CLS token and a 4×4 grid of visual tokens per frame to an 18-layer, width-1,024 DiT with 16 attention heads. The policy also uses current-image score features from its internal predictor heads. It predicts a chunk of 32 camera-local increments, executes the first, and replans after receiving the resulting observation.

Each increment contains three translation coordinates and the first two columns of a relative rotation; a compatibility channel is excluded from action supervision. Stored per-channel quantiles $q_{.01,j}$ and $q_{.99,j}$ define the inverse normalization:

$$a_j = \frac{\hat{z}_j + 1}{2}(q_{.99,j} - q_{.01,j}) + q_{.01,j}. \tag{6}$$

The decoded rotation and translation pass through the same geometric executor as the baselines.

Training with Internet, game, and 3DGS data. Internet videos with estimated motion and game/3DGS demonstrations with recorded poses supply causal windows in the same framework. Inputs contain past observations and executed actions; future increments are targets, with missing steps masked. Training optimizes the DiT, visual projections, and final 16 ViT-L blocks; internal scoring heads remain frozen.

For a normalized action target z_0, training samples $z_t = (1-t)z_0 + t\epsilon$, with $t \sim U(0,1)$ and Gaussian noise ϵ. The model predicts velocity v_θ using the masked flow-matching loss

$$\mathcal{L} = \mathbb{E}\left[\frac{\|M \odot (v_\theta(z_t, t) - (\epsilon - z_0))\|_2^2}{\|M\|_1}\right], \tag{7}$$

where M selects valid future translation and rotation channels. Eight decoding steps produce the action chunk. Executed motion, including rejections, updates the history; every request consumes interaction budget.

7 Experiments

We evaluate trajectory supervision, camera execution, and policy performance over 49 scenes. Figure 7 summarizes the training and deployment of BSCP (Section 6).

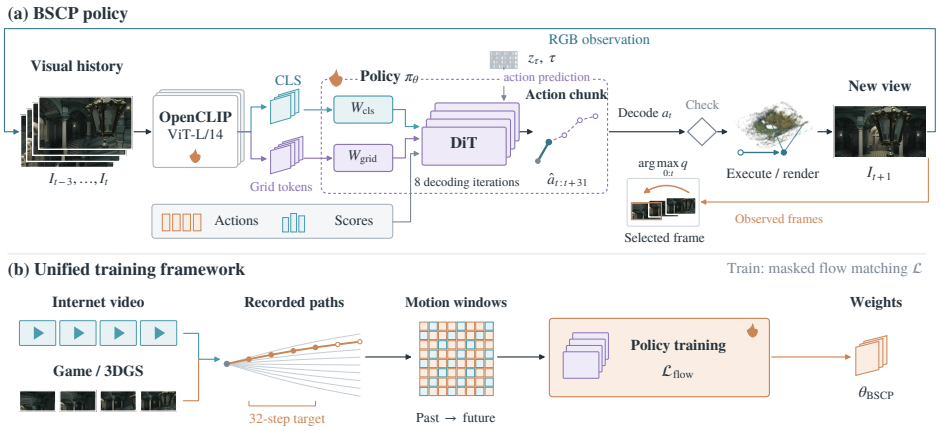

Figure 7: **BSCP within EZCAM.** (a) History conditions 32 predicted increments; the first is executed before replanning. (b) Internet and game/3DGS data supply motion targets. Images, paths, and source proportions are schematic.

7.1 Data and execution evidence

Sections 4–5 report data and execution checks (Table 15). Development-service render calls take 143, 140, and 198 ms median for Wukong, KCD2, and campus, including transport and queueing.

7.2 Closed-loop comparison over 49 scenes

Population and protocol. Six starts per asset give 294 episodes per model. PF41 fixes 41 observations and 40 requests without future-view previews. The split contains ten game/3DGS training, three development, and 36 unseen scenes. Starts use the safety-volume FPS6 manifest: pilot poses

are preserved; other orientations follow published cameras or object centers, without aesthetic selection.

Trained models and adaptation. The main table compares seven EZCAM-trained models under a shared execution and evaluation protocol, using the same training datasets, which consist of game-based data, 3DGS data and Internet-sourced data. Greedy search over unexecuted candidate views violates PF41's zero-preview setting; executed search pays for each move and observation. A separate timing test measured scoring latency at about $70\times$ one policy action prediction. Random local and Forward obtain mean gains of 0.2186 and 0.1096, below BSCP's 0.5126.

Table 1: **Trained-model comparison over 49 scenes.** Six shared starts per scene, 40 requests. Splits follow game/3DGS training, development, and scenes unseen in these stages. Baselines are trained feature adapters; $SR_{0.3}$ is in percent; bold marks BSCP.

	All 49				Seen (10)	Dev. (3)	Unseen (36)	
Model	G_{best}	AUC	$SR_{0.3}$	G_{end}	G_{best}	G_{best}	G_{best}	$SR_{0.3}$
BSCP (ours)	**0.5126**	**0.3857**	**57.8**	**-0.3653**	**0.8480**	**0.7638**	**0.3985**	**47.2**
DualVLN	0.2817	0.2438	30.6	-0.4196	0.7919	0.6483	0.1094	13.9
StreamVLN	0.2637	0.2159	32.0	-0.3563	0.4167	0.4538	0.2054	23.6
Uni-NaVid	0.2302	0.1861	31.6	-0.2227	0.4087	0.4755	0.1602	21.3
NavGPT-2	0.2171	0.1841	27.6	-0.3392	0.5416	0.5425	0.0999	12.0
NaviLLM	0.2077	0.1755	27.9	-0.4049	0.3703	0.5440	0.1345	15.7
NaVid	0.2006	0.1679	23.5	-0.2561	0.3373	0.4895	0.1386	13.9

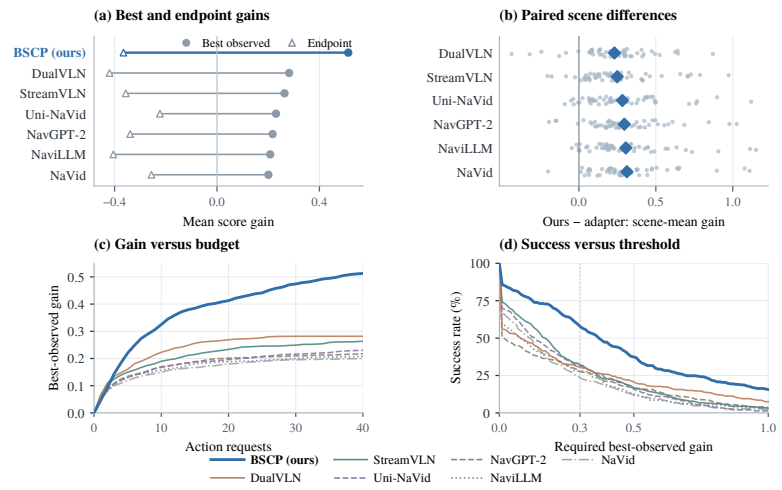

Figure 8: **Trained-model comparison.** (a) Best/endpoint gains. (b) BSCP minus adapters: 49 scene means and their mean (diamonds), descriptive rather than significance tests. (c) Gain versus budget. (d) Success versus threshold. Same 294 episodes per model as Table 1.

Aggregate and unseen-scene results. BSCP leads the seven models in mean gain (0.5126), AUC (0.3857), and $SR_{0.3}$ (57.8%); DualVLN reaches 0.2817 gain. The paired difference is $+0.2309$, with 43 scene wins and six losses (Figure 8). On 36 scenes excluded from game/3DGS training and development, BSCP reaches 0.3985 gain, 47.2% success, and 0.3006 AUC, versus 0.2054, 23.6%, and 0.1776 for StreamVLN, the strongest adapter on this subset. This supports transfer beyond the adaptation scenes.

Budget and photographic output. Mean best-observed gain reaches 0.3249 at 10 requests and 0.4126 at 20 (Figure 8c); panel (d) varies the success threshold. Endpoint gain is -0.3653, with 62.9% ending below the initial score: retaining a photograph does not imply stopping or returning at its pose. Figure 9 illustrates selected photographs from four matched starts using the same checkpoint as the main table. All 2,058 complete score sequences reproduce the metrics; starting poses, seeds, scores, and RGB hashes agree across models (Appendix E).

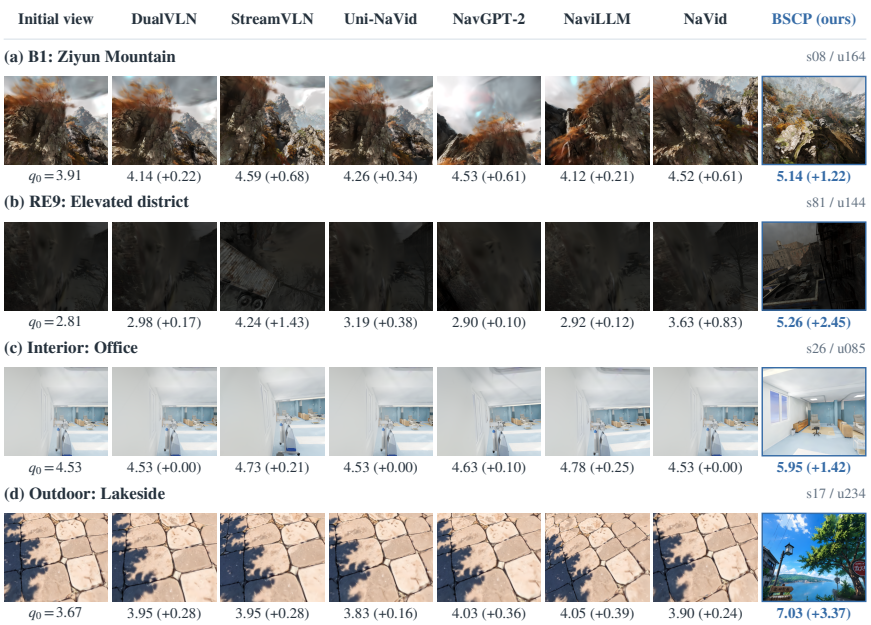

Figure 9: **Best-frame selections from identical starts.** Highest-scoring frames among 41 observations; labels show score q and gain $q - q_0$. Examples favor scene variety, visual contrast, and larger margins. Source pixels (640×544) are unchanged; adapters match Table 1.

7.3 Ablations

Training data. We treat game and 3DGS trajectories as one bundle and compare three data sources under the same policy configuration (Table 2). Mean best-observed gain is 0.4453 with the bundle alone, 0.4169 with Internet data alone, and 0.5126 with both (also Table 3, final row).

Reference controls. Table 3 lists uniform and branch-supervised reference policies alongside full-data BSCP under the shared 294-start, 40-request evaluation. Matched supervision/recovery contrasts are in Appendix E.3.

Table 2: **Data ablation.**

Game/3DGS	Internet	$G_{\text{best}} \uparrow$
✓	—	0.4453
—	✓	0.4169
✓	✓	**0.5126**

Table 3: **Reference variants and full-data BSCP.**

Variant	Recovery	$G_{\text{best}} \uparrow$	AUC↑	$\text{SR}_{0.3}(\%)$	G_{end}
Uniform	On	0.4022	0.3208	49.0	−0.3014
BSCP	Off	0.3965	0.3367	39.1	−0.4289
BSCP	On	**0.5126**	**0.3857**	**57.8**	**-0.3653**

BSCP achieves 0.5126 gain, 0.3857 AUC, and 57.8% success.

8 Limitations

EZCAM currently studies camera policies in static reconstructed environments. Extending to dynamic settings and broader scene domains is a natural next step. Comparisons under matched training budgets would further clarify the contributions of policy architecture and supervision.

9 Conclusion

EZCAM connects photographic supervision, executable camera interaction, and a policy trained with Internet, game, and 3DGS data. Complete-path checks reject requests missed by endpoint-only validation. BSCP leads seven trained models in best-observed gain, AUC, and success over 49 scenes, while mean endpoint gain remains negative. Controls show higher mean return with branch supervision and recovery. Retained-image and endpoint quality remain distinct.

REPRODUCIBILITY STATEMENT

The supplement includes the PF41 protocol, shared start poses, all 2,058 score sequences for the seven-model comparison, and the scene-split manifest. Checkpoint identifiers, source hashes, and aggregation scripts reproduce the main table and comparison curves. The selected RGB frames retain their recorded pixels, scores, and poses. The 882 controlled ablation records and paired summaries reproduce the ablation table. Geometry checks, render-call records, 2,220 anonymized human responses, and their analysis scripts are also supplied. Reconstruction records include 128 KCD2 source/render pairs and the archived RE9 fidelity report. Public backbones and source media remain separately identified dependencies.

ETHICS STATEMENT

Aesthetic preferences vary across viewers and cultures; predictor agreement is not a universal definition of quality. Source videos and scenes may contain people, identifying details, or copyrighted assets. All 37 study participants provided informed consent. We distribute anonymous questionnaire identifiers and responses, excluding private respondent identifiers. The authors report that the applicable source agreements permit noncommercial research use; source media, game assets, and pretrained models remain subject to their respective terms. This statement does not grant unrestricted redistribution rights. Physical deployment requires safeguards beyond point-camera simulation.

AI USE STATEMENT

AI assistance was used for manuscript editing, evidence organization, and verification scripts. No additional AI use is reported. Reported policy metrics and human judgments come from the archived experiments and study; AI assistance did not supply substitute judgments or experimental outcomes. The authors are responsible for the manuscript and its claims.

REFERENCES

Hadi AlZayer, Hubert Lin, and Kavita Bala. AutoPhoto: Aesthetic photo capture using reinforcement learning. In *IROS*, 2021. doi: 10.1109/IROS51168.2021.9636788. URL https://arxiv.org/abs/2109.09923.

Peter Anderson, Qi Wu, Damien Teney, Jake Bruce, Mark Johnson, Niko Sünderhauf, Ian Reid, Stephen Gould, and Anton van den Hengel. Vision-and-language navigation: Interpreting visually-grounded navigation instructions in real environments. In *CVPR*, 2018.

Ruzena Bajcsy. Active perception. *Proceedings of the IEEE*, 76(8), 1988.

Jonathan T. Barron, Ben Mildenhall, Dor Verbin, Pratul P. Srinivasan, and Peter Hedman. Mip-NeRF 360: Unbounded anti-aliased neural radiance fields. In *CVPR*, 2022.

Kevin Black, Noah Brown, Danny Driess, Adnan Esmail, Michael Equi, Chelsea Finn, et al. π_0: A vision-language-action flow model for general robot control. *arXiv preprint*, 2024.

Anthony Brohan, Noah Brown, Justice Carbajal, Yevgen Chebotar, Xi Chen, Krzysztof Choromanski, et al. RT-2: Vision-language-action models transfer web knowledge to robotic control. In *Conference on Robot Learning (CoRL)*, 2023.

Shuo Cao, Nan Ma, Jiayang Li, Xiaohui Li, Lihao Shao, Kaiwen Zhu, Yu Zhou, Yuandong Pu, Jiarui Wu, Jiaquan Wang, Bo Qu, Wenhai Wang, Yu Qiao, Dajuin Yao, and Yihao Liu. ArtiMuse: Fine-grained image aesthetics assessment with joint scoring and expert-level understanding. *arXiv preprint arXiv:2507.14533*, 2025. URL https://arxiv.org/abs/2507.14533.

C. Connolly. The determination of next best views. In *IEEE International Conference on Robotics and Automation (ICRA)*, 1985.

Robin Courant, Nicolas Dufour, Xi Wang, Marc Christie, and Vicky Kalogeiton. E.T. the exceptional trajectories: Text-to-camera-trajectory generation with character awareness. In *ECCV*, 2024. URL https://arxiv.org/abs/2407.01516.

discus0434. Aesthetic Predictor V2.5. GitHub repository, 2024. URL https://github.com/discus0434/aesthetic-predictor-v2-5. Accessed September 18, 2026.

Peter Hedman, Julien Philip, True Price, Jan-Michael Frahm, George Drettakis, and Gabriel Brostow. Deep blending for free-viewpoint image-based rendering. *ACM Transactions on Graphics*, 37(6), 2018.

Hongda Jiang, Xi Wang, Marc Christie, Libin Liu, and Baoquan Chen. Cinematographic camera diffusion model. *Computer Graphics Forum*, 2024. doi: 10.1111/cgf.15055. URL https://arxiv.org/abs/2402.16143.

Bernhard Kerbl, Georgios Kopanas, Thomas Leimkühler, and George Drettakis. 3D gaussian splatting for real-time radiance field rendering. *ACM Transactions on Graphics*, 42(4), 2023.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, Quan Vuong, Thomas Kollar, Benjamin Burchfiel, Russ Tedrake, Dorsa Sadigh, Sergey Levine, Percy Liang, and Chelsea Finn. OpenVLA: An open-source vision-language-action model. In *Conference on Robot Learning (CoRL)*, 2024.

Arno Knapitsch, Jaesik Park, Qian-Yi Zhou, and Vladlen Koltun. Tanks and temples: Benchmarking large-scale scene reconstruction. *ACM Transactions on Graphics*, 36(4), 2017.

Shu Kong, Xiaohui Shen, Zhe Lin, Radomir Mech, and Charless Fowlkes. Photo aesthetics ranking network with attributes and content adaptation. In *ECCV*, 2016.

Xinyang Li, Zhangyu Lai, Linning Xu, Yansong Qu, Liujuan Cao, Shengchuan Zhang, Bo Dai, and Rongrong Ji. Director3D: Real-world camera trajectory and 3D scene generation from text. In *NeurIPS*, 2024.

Xinhang Liu, Yu-Wing Tai, and Chi-Keung Tang. ChatCam: Empowering camera control through conversational AI. In *NeurIPS*, 2024. URL https://arxiv.org/abs/2409.17331.

Ben Mildenhall, Pratul P. Srinivasan, Matthew Tancik, Jonathan T. Barron, Ravi Ramamoorthi, and Ren Ng. NeRF: Representing scenes as neural radiance fields for view synthesis. In *ECCV*, 2020.

Naila Murray, Luca Marchesotti, and Florent Perronnin. AVA: A large-scale database for aesthetic visual analysis. In *CVPR*, 2012.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, 2021.

Manolis Savva, Abhishek Kadian, Oleksandr Maksymets, Yili Zhao, Erik Wijmans, Bhavana Jain, Julian Straub, Jia Liu, Vladlen Koltun, Jitendra Malik, Devi Parikh, and Dhruv Batra. Habitat: A platform for embodied AI research. In *ICCV*, 2019.

Christoph Schuhmann. LAION-Aesthetics. LAION project blog, August 16, 2022. URL https://laion.ai/blog/laion-aesthetics/.

Hossein Talebi and Peyman Milanfar. NIMA: Neural image assessment. *IEEE Transactions on Image Processing*, 27(8), 2018.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Yi Zhang, Chang Zeng, Yanxi Zhou, Xiaoxiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. SpatialVID: A large-scale video dataset with spatial annotations. *arXiv preprint arXiv:2509.09676*, 2025. URL https://arxiv.org/abs/2509.09676.

Meng Wei, Chenyang Wan, Jiaqi Peng, Xiqian Yu, Yuqiang Yang, Delin Feng, Wenzhe Cai, Chenming Zhu, Tai Wang, Jiangmiao Pang, and Xihui Liu. Ground slow, move fast: A dual-system foundation model for generalizable vision-and-language navigation. *arXiv preprint arXiv:2512.08186*, 2025a. URL https://arxiv.org/abs/2512.08186.

Meng Wei, Chenyang Wan, Xiqian Yu, Tai Wang, Yuqiang Yang, Xiaohan Mao, Chenming Zhu, Wenzhe Cai, Hanqing Wang, Yilun Chen, Xihui Liu, and Jiangmiao Pang. StreamVLN: Streaming vision-and-language navigation via slowfast context modeling. *arXiv preprint arXiv:2507.05240*, 2025b. URL https://arxiv.org/abs/2507.05240.

Zijun Wei, Jianming Zhang, Xiaohui Shen, Zhe Lin, Radomír Měch, Minh Hoai, and Dimitris Samaras. Good view hunting: Learning photo composition from dense view pairs. In *CVPR*, 2018.

Lujian Yao, Siming Zheng, Xinbin Yuan, Zhuoxuan Cai, Pu Wu, Jinwei Chen, Bo Li, and Peng-Tao Jiang. Photography perspective composition: Towards aesthetic perspective recommendation. In *NeurIPS*, 2025. URL https://papers.neurips.cc/paper_files/paper/2025/hash/8227285e32f70e07fa3a247f3a48006d-Abstract-Conference.html.

Zhongrui You, Zhen Li, Junli Liu, Zhigang Wang, and Bin Zhao. OpenFlyScan: A quality-guided aerial reconstruction system for consumer drones, 2026. URL https://arxiv.org/abs/2609.24253.

Hui Zeng, Lida Li, Zisheng Cao, and Lei Zhang. Reliable and efficient image cropping: A grid anchor based approach. In *CVPR*, 2019.

Rui Zeng, Yuhui Wen, Wang Zhao, and Yong-Jin Liu. View planning in robot active vision: A survey of systems, algorithms, and applications. *Computational Visual Media*, 6(3):225–245, 2020.

Jiazhao Zhang, Kunyu Wang, Rongtao Xu, Gengze Zhou, Yicong Hong, Xiaomeng Fang, Qi Wu, Zhizheng Zhang, and He Wang. NaVid: Video-based VLM plans the next step for vision-and-language navigation. In *Robotics: Science and Systems (RSS)*, 2024.

Jiazhao Zhang, Kunyu Wang, Shaoan Wang, Minghan Li, Haoran Liu, Songlin Wei, Zhongyuan Wang, Zhizheng Zhang, and He Wang. Uni-NaVid: A video-based vision-language-action model for unifying embodied navigation tasks. In *Robotics: Science and Systems (RSS)*, 2025a.

Mengchen Zhang, Tong Wu, Jing Tan, Ziwei Liu, Gordon Wetzstein, and Dahua Lin. GenDoP: Auto-regressive camera trajectory generation as a director of photography. In *ICCV*, pp. 18229–18239, 2025b. URL https://openaccess.thecvf.com/content/ICCV2025/html/Zhang_GenDoP_Auto-regressive_Camera_Trajectory_Generation_as_a_Director_of_Photography_ICCV_2025_paper.html.

Duo Zheng, Shijia Huang, Lin Zhao, Yiwu Zhong, and Liwei Wang. Towards learning a generalist model for embodied navigation. In *CVPR*, 2024.

Gengze Zhou, Yicong Hong, Zun Wang, Xin Eric Wang, and Qi Wu. NavGPT-2: Unleashing navigational reasoning capability for large vision-language models. *arXiv preprint arXiv:2407.12366*, 2024. URL https://arxiv.org/abs/2407.12366.