NEXTSHOT: LEARNING SOURCE-GROUNDED STRUCTURED COMPOSITION TARGETS FOR PORTRAIT PHOTOGRAPHY

Anonymous authors
Paper under double-blind review

ABSTRACT

Modern visual models increasingly move beyond recognizing isolated content toward structured understanding of scene entities, semantics, and relationships. Relational information is useful not only for characterizing what is observed: when a task requires inferring a desired configuration from existing visual content, that configuration often depends on interactions among entities and their global context. We study this problem in photographic composition, where relevant scene elements are already present and composition depends on how people and environmental regions are positioned and scaled relative to one another and the frame. We formulate **Region-Level Composition Target Prediction**, which preserves source-region identities while jointly predicting regional geometry, target frame aspect ratio, and person framing. Because photographic demonstrations typically provide only whole-image source–reference outcomes, direct region-level supervision is unavailable. We recover structured supervision by transferring reference geometry to source identities through partial correspondence while preserving each reference as a complete alternative arrangement, complemented by single-image aesthetic annotations. Our predictor, **NextShot**, models region–region and region–frame relations, integrates multimodal scene semantics, and conditions regional geometry on predicted frame and person framing. NextShot improves centroid and visible-area prediction over an autoregressive alternative. Two independent large multimodal evaluators from different model families also consistently prefer its complete layouts over alternative predictors and key model variants. In a 100-case guided-recapture study, five independent human evaluators prefer photographs taken with NextShot guidance in **78.0% of 500 judgments**. These results establish source-grounded structured targets as a learnable representation for explicit spatial goal prediction from relational scene information.

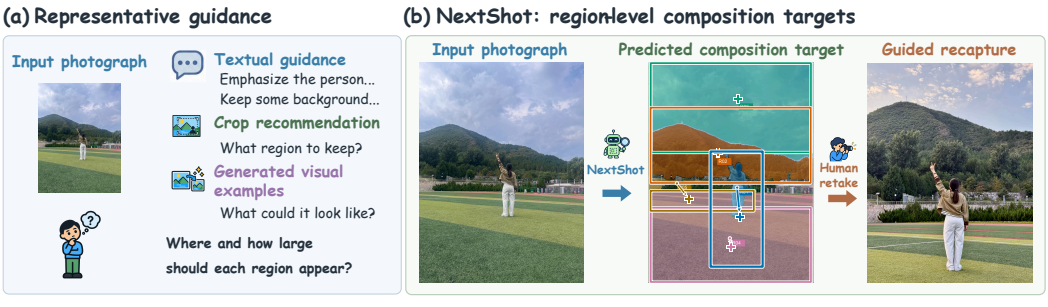

Figure 1: From photographic feedback to spatial targets. (a) Existing guidance offers critiques, crops, or visual exemplars. (b) NextShot specifies where the observed regions should appear and how large they should be in the next photograph, providing spatial targets for guided recapture.

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.

1 INTRODUCTION

Structured visual understanding represents scenes through entities, attributes, and relationships (Mitra et al., 2024; Park et al., 2025; Yang et al., 2026). Semantic and spatial constraints also guide layout generation (Feng et al., 2023; Wang et al., 2025). We study this connection in photographic composition: predicting a desired configuration of observed regions, where each region's target state depends on the others and their shared context.

In portrait photography, the desired positions, scales, and framing of people and environmental regions are interdependent. Enlarging a person changes the prominence of the surroundings, while changing the frame alters every region's spatial context. Photographic reconstruction and generated exemplars visualize improved outcomes (Du et al., 2026a; You et al., 2026). Our aim is to make the regional configuration underlying such outcomes an explicit, source-grounded learning target.

We formulate **Region-Level Composition Target Prediction**. Given a portrait photograph, the task jointly predicts regional geometry, target frame aspect ratio, and coarse person framing. Source-region identities anchor each prediction to observed content, while target framing provides a shared context for regional geometry. The resulting representation supports learning and evaluating complete arrangements, and gives photographers an explicit spatial goal for recapture (Figure 1).

Learning this target requires regional supervision that whole-image photographic demonstrations do not directly provide. Across a source–reference pair, regions can change position, scale, and visibility; multiple references may also express distinct valid compositions. Recovering targets therefore requires associating regions while preserving the coherence of each demonstrated arrangement. We construct **AesRecon-CT** from AesRecon (Du et al., 2026a) by transferring reference geometry through partial source–reference correspondence. Each reference remains a complete alternative, with its valid regional targets supervised jointly.

A region's desired placement and scale depend on its semantic role and relationships to the surrounding scene. Target frame and person framing further specify the canvas and intended visible extent for these predictions. Our predictor, **NextShot**, models these dependencies by encoding region–region and region–frame relations, integrating multimodal context, and conditioning regional geometry on predicted frame and person framing. We complement geometric demonstrations with **PortraitCraft-Aux**, using attribute scores and rationales from PortraitCraft (Sha et al., 2026) to train the shared representation.

We evaluate the targets through geometric agreement, complete-layout preference, and guided recapture. NextShot improves centroid and visible-area prediction over an autoregressive alternative. Two multimodal evaluators from different model families prefer its layouts over this alternative and variants without aesthetic supervision, explicit relational biases, or framing conditioning. In a 100-case guided-recapture study, five independent human evaluators prefer guided photographs in **78.0% of 500 judgments**.

Our contributions are:

- A source-grounded structured target formulation for photographic composition, jointly representing regional geometry, target frame, and person framing.
- Structured supervision recovered from whole-image demonstrations through partial correspondence and complete-reference training, complemented by aesthetic annotations.
- A relational, multimodal, and framing-conditioned predictor, evaluated through geometric agreement, layout preference, and real guided recapture.

2 RELATED WORK

Photographic composition and guidance. Cropping methods select image frames using photographic examples, spatial–semantic relations, or vision–language context (Hong et al., 2024; Zhang et al., 2026; Su et al., 2024; Lee et al., 2025). Viewpoint methods predict camera adjustments from a single image or optimize views in a 3D aesthetic field built from sparse observations (Li et al., 2025; Tang et al., 2026), while multimodal assistants provide aesthetic and pose guidance (Du et al., 2026b; Li et al., 2026). PPC demonstrates perspective improvements through generated videos (Yao

et al., 2025); PhotoFramer and AesFormer visualize outcomes through instruction–exemplar pairs and structural reconstruction (You et al., 2026; Du et al., 2026a). NextShot formulates the desired arrangement of source-associated people and environmental regions as a structured prediction target.

Structured layout and relational modeling. Structured layout prediction connects element content, spatial constraints, and global context. LayoutFormer++ handles explicit layout constraints (Jiang et al., 2023), while LaDeCo arranges supplied multimodal elements through semantic layer planning and conditioning on previously rendered layers (Lin et al., 2025). AesthetiQ further aligns predicted layouts with aesthetic preferences (Patnaik et al., 2025). Scene-graph methods use entity relations for compositional visual reasoning through intermediate prompts or graph-derived training signals (Mitra et al., 2024; Yang et al., 2026). Our setting concerns the visible geometry of people and extended environmental regions observed in the source photograph. Their target geometry is coupled to photographic framing, and region-level supervision is recovered through source–reference correspondence.

Aesthetic understanding and supervision. Multimodal aesthetic research spans visual scoring (Wu et al., 2024), photographic critique (Qi et al., 2025), and fine-grained attribute analysis (Cao et al., 2026). PICD benchmarks composition understanding (Zhao et al., 2025), while PortraitCraft supports portrait-composition understanding and generation with scores, attributes, and rationales (Sha et al., 2026). AesNet learns aesthetic representations through attribute-level image–text alignment (Huang et al., 2025). We use PortraitCraft's attribute scores and rationales as auxiliary supervision for target prediction.

3 Structured Composition Targets and Supervision

3.1 Region-Level Composition Targets

Given a portrait photograph I and composition-relevant source regions $\mathcal{R} = \{r_i\}_{i=1}^{N}$, we predict a two-dimensional target retaining their identities:

$$\mathcal{T} = \left(\rho,\ \{\,\mathbf{b}_i, \mathbf{c}_i, a_i\,\}_{i=1}^{N},\ \{\,\mathbf{f}_i\,\}_{i\in\mathcal{P}}\right),$$

where ρ is the target log aspect ratio and $\mathcal{P}$ indexes person regions. Each region has a target visible-region box $\mathbf{b}_i$, mask-derived centroid $\mathbf{c}_i$, and visible-area fraction a_i, normalized by the target frame. For people, $\mathbf{f}_i$ is a coarse distribution over full-body and half-body presentation.

Boxes describe spatial extent, centroids locate visible content, and areas measure its share of the frame. These fields capture different properties of irregular environmental regions that only partly occupy their boxes. The frame defines their shared canvas, while person framing specifies intended visible extent.

3.2 Constructing Source-Grounded Targets

We construct **AesRecon-CT** from AesRecon (Du et al., 2026a) by transferring reference geometry to source-region identities (Figure 2).

Photographic and regional evidence. Each group contains a source and one or more higher-quality references from the same photographic event. After curation for portrait relevance and usable content, pretrained models identify, localize, and segment people and environmental regions. Masks supply geometry; appearance and semantics support correspondence, while available body, orientation, and depth evidence also inform prediction. Appendices A.1 and A.3 detail this processing.

Partial correspondence. Independently extracted source and reference regions are associated by partial bipartite assignment using appearance and semantic compatibility. Position and scale are excluded because composition changes are expected to modify them. Matches must be supported relative to competing assignments; unmatched regions are permitted.

Accepted matches transfer reference boxes, visible centroids, and areas to source identities in the reference frame. Ambiguous or unmatched source regions remain context and prediction variables

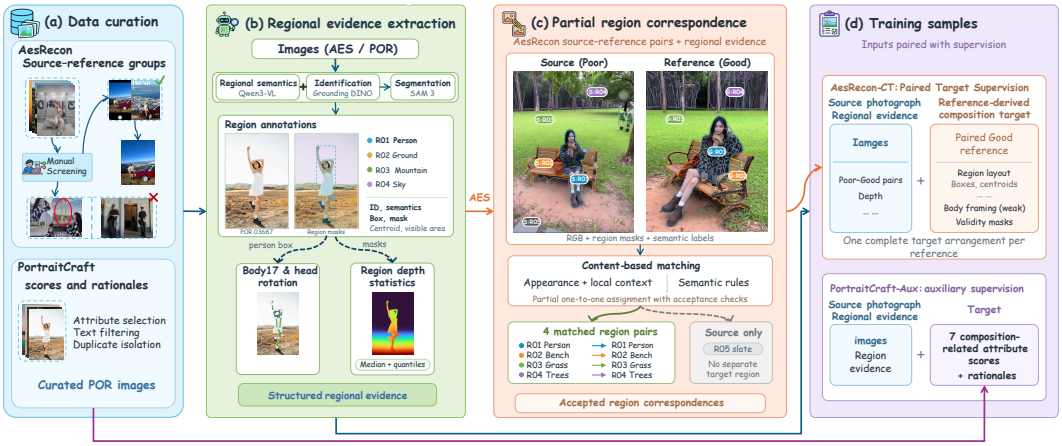

Figure 2: Source-grounded training supervision. After curation and regional evidence extraction, source–reference regions are partially matched without position or scale cues. **AesRecon-CT** supplies paired geometry and framing targets; **PortraitCraft-Aux** supplies aesthetic attribute scores and rationales.

without geometric supervision from that reference; reference-only regions add no output identity. Inference uses only the source photograph and its extracted evidence.

Complete-reference supervision. Each reference is retained as a complete alternative arrangement, with its own valid correspondences and choices of subject scale, environment, and framing. Joint supervision within one reference preserves these coordinated choices (Section 4.4); mixing regions across references or averaging alternatives could produce an undemonstrated arrangement. Reference-body evidence supplies weak full-/half-body labels for matched people; unresolved cases remain unsupervised.

Auxiliary aesthetic annotations. **PortraitCraft-Aux** supplies scores and rationales for seven composition-related attributes from PortraitCraft (Sha et al., 2026) (Appendix A.4). These annotations describe observed photographs and supervise the shared representation; AesRecon-CT supplies the desired geometry and framing.

4 NextShot

NextShot maps source regions to the target $\mathcal{T}$ through relational encoding, multimodal fusion, and framing-conditioned decoding (Figure 3).

4.1 Region–Frame Relational Encoding

We encode the source arrangement with one node per region r_i and a frame anchor r_F, jointly modeling region–region and region–boundary relations. Region inputs combine a frozen appearance descriptor, source geometry (box, centroid, area, and available depth statistics), and available orientation evidence. The frame anchor combines a learned embedding with source-canvas geometry. A learned encoder maps these inputs to states $\mathbf{h}_i^0$ and $\mathbf{h}_F^0$.

For each node pair, a source-derived vector $\mathbf{e}_{ij}$ describes relative centroid displacement, visible scale, centroid distance, box overlap, normalized overlap geometry, and semantic compatibility. A small network maps these features to a head-specific attention bias:

$$\ell_{ij}^{(m)} = \frac{(\mathbf{W}_q^{(m)}\mathbf{h}_i)^\top(\mathbf{W}_k^{(m)}\mathbf{h}_j)}{\sqrt{d_m}} + \beta_m(\mathbf{e}_{ij}),$$

where m indexes heads, d_m is the query/key dimension, and β_m is the learned bias function. Softmax over valid nodes j, followed by attention and feed-forward updates, yields states $\mathbf{h}_i$ and $\mathbf{h}_F$.

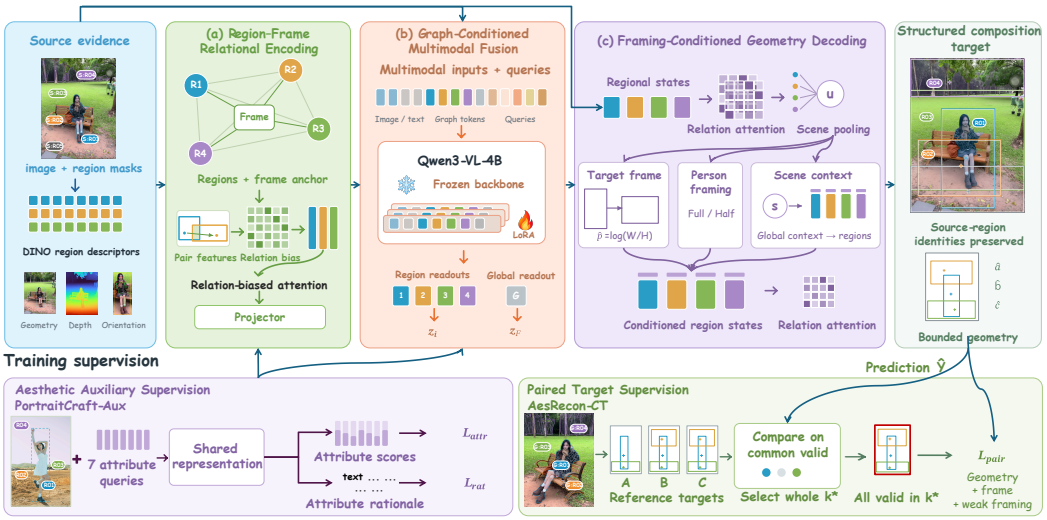

Figure 3: NextShot architecture. Region–region and region–frame relations combine with multimodal context to yield region-aligned and global readouts. The decoder first predicts the target frame and person framing, then conditions regional geometry on them. Training combines paired targets from AesRecon-CT with auxiliary aesthetic supervision from PortraitCraft-Aux.

4.2 Graph-Conditioned Multimodal Fusion

Projected regional and frame states enter a vision–language backbone as continuous graph tokens alongside the source image and textual region descriptions. To retain a dedicated readout for each source identity, we initialize regional and global queries as

$$\mathbf{q}_i^0 = P(\mathbf{h}_i) + \mathbf{q}_{\text{reg}}, \qquad \mathbf{q}_F^0 = P(\mathbf{h}_F) + \mathbf{q}_{\text{global}}.$$

Here P is a shared nonlinear projection, and $\mathbf{q}_{\text{reg}}$ and $\mathbf{q}_{\text{global}}$ are learned embeddings. Multimodal processing yields readouts $\mathbf{z}_i$ and $\mathbf{z}_F$ for numerical decoding. Each regional query accesses scene context while retaining its source association.

We use Qwen3-VL-4B-Instruct (Bai et al., 2025) with frozen base weights. The relational encoder, graph projector, query embeddings, numerical decoder, and LoRA (Hu et al., 2022) adapters in the language layers are trained jointly.

4.3 Framing-Conditioned Geometry Decoding

To align regional geometry with the intended canvas and visible person extent, the decoder predicts frame and person framing first. It fuses regional readouts $\mathbf{z}_i$, global readout $\mathbf{z}_F$, source states $\mathbf{h}_i$, geometry, and available body evidence. Relational updates yield $\mathbf{s}_i$; attention pooling and fusion with $\mathbf{z}_F$ produce the scene state $\mathbf{s}$.

With source log aspect ratio ρ_s, the frame and person-framing heads predict

$$\hat{\rho} = \rho_s + \Delta\rho(\mathbf{s}), \qquad \hat{\mathbf{f}}_i = \text{softmax}\big(g_f(\mathbf{s}_i, \mathbf{s})\big),$$

where $\hat{\mathbf{f}}_i$ gives full-/half-body probabilities for person regions. We inject these predictions before jointly updating all regions:

$$\mathbf{v}_i = \mathbf{s}_i + \mathbf{s} + E_\rho(\hat{\rho}) + \mathbf{t}_i,$$
$$\{\tilde{\mathbf{s}}_i\}_{i=1}^N = \text{RelBlock}\big(\{\mathbf{v}_i\}_{i=1}^N, \{\mathbf{e}_{ij}\}\big).$$

Here $\mathbf{t}_i = E_f(\hat{\mathbf{f}}_i)$ for person regions and $\mathbf{t}_i = 0$ otherwise. RelBlock propagates framing information across regions using the source relations; numerical heads then predict boxes, visible centroids, and visible-area fractions.

Bounded residual adjustments to source geometry keep boxes inside the target frame, centroids inside their boxes, and visible-area fractions within the enclosing box area.

4.4 LEARNING OBJECTIVES

Paired targets. Let $\{\mathcal{T}^{(k)}\}_{k=1}^{K}$ be a source's references, V_k the valid regions in reference k, and C the region set valid across the available references. Candidate references are compared on C:

$$D_k = D_{\text{reg}}\left(\hat{\mathcal{T}}, \mathcal{T}^{(k)}; C\right) + \lambda_\rho D_\rho\left(\hat{\rho}, \rho^{(k)}\right).$$

Here D_{reg} combines smooth-L_1 errors on boxes, visible centroids, and log visible areas. Errors are averaged within person and environment groups, then across nonempty groups, balancing their contributions. The common set C makes reference scores comparable despite differing correspondence coverage.

We select $k^* = \arg\min_k D_k$ and apply the same geometric errors to all valid regions in V_{k^*}. Selection thus compares shared evidence, while training uses all supported regions within the selected arrangement. The paired loss is

$$\mathcal{L}_{\text{pair}} = \mathcal{L}_{\text{geom}} + \lambda_\rho \mathcal{L}_{\text{frame}} + \lambda_f \mathcal{L}_{\text{framing}},$$

with cross-entropy for available weak full-/half-body labels.

Aesthetic supervision. For each PortraitCraft-Aux image, seven attribute queries read the shared multimodal representation. Available normalized scores use masked smooth-L_1 regression, $\mathcal{L}_{\text{attr}}$; one available rationale uses answer-token cross-entropy, $\mathcal{L}_{\text{rat}}$. Attribute readouts precede the rationale answer in the causal sequence, so score prediction does not condition on a teacher-forced explanation.

The full objective is

$$\mathcal{L} = \mathcal{L}_{\text{pair}} + \lambda_{\text{attr}} \mathcal{L}_{\text{attr}} + \lambda_{\text{rat}} \mathcal{L}_{\text{rat}}.$$

Auxiliary gradients update the shared relational encoder, graph projector, and backbone adapters; only paired supervision trains the composition decoder. Appendix A.5 gives loss scales and training settings. At inference, auxiliary heads are removed and one pass predicts the target.

5 EXPERIMENTS

We evaluate target prediction through reference agreement and layout preference, examine model design, and assess guided recapture in practice.

5.1 EXPERIMENTAL SETUP

Data and evaluation sets. Paired training uses 2,915 source photographs with 4,268 valid references, supplemented by 25,164 PortraitCraft-Aux images. The 450-source development set supports checkpoint selection and controlled comparisons and emphasizes visible composition changes. Reference agreement uses 781 sources with eligible geometric references from the retained AesRecon test assignment. Layout evaluation uses fixed random samples: 200 test sources shared by both evaluators, and 200 development sources for Qwen with a random 100-source subset for Flash. The photography study uses 100 pairs collected independently of AesRecon-CT. Splits keep source groups intact; Appendix A.2 details selection and overlap checks.

Compared systems. NoChange preserves source geometry. NoVLM-Paired is a lightweight structured predictor trained on paired targets; NoVLM-AttrAux adds attribute supervision without rationales. The 4B w/o AesAux variant uses the NextShot architecture with paired supervision alone. The w/o RelBias variant removes explicit attention biases, retaining ordinary attention and the remaining architecture. The w/o Framing Conditioning variant removes predicted frame and person-framing inputs to geometry decoding, retaining their heads, relational layers, and losses. NextShot-8B scales the backbone under the joint objective (Appendix B.2). AR-8B adapts an 8B VLM to autoregressively predict centroids and visible areas; ZS-32B uses a 32B VLM without task-specific training.

Evaluation. Reference metrics are box IoU, visible-centroid and visible-area MAE, and frame aspect-ratio relative error (RE). Regional metrics are averaged within sources and then across

sources, with separate person and environment IoUs. Reference agreement measures fit to a demonstrated composition; layout preference assesses the complete arrangement. Qwen and Flash compare subject hierarchy, spatial balance, subject–environment relations, and overall expression. Preference rates use all planned judgments as the denominator (Appendix F).

5.2 Main Target Prediction

Table 1: Main test-set results. Reference agreement uses 781 sources; layout preference uses 200 shared sources. Q/F denote Qwen/Flash, with win rates measured for NextShot over the row method. AR-8B predicts only centroids and visible areas.

	Reference agreement						Layout preference	
	Box IoU ↑			MAE ↓		Frame RE% ↓	NextShot win% vs. row	
Method	All	Person	Env.	Centroid	Area		Q	F
Baselines								
NoChange	0.5639	0.4914	0.6079	0.0745	0.0781	9.848	89.5	77.0
AR-8B	—	—	—	0.0680	0.0727	—	79.5	81.0
Core variants								
NoVLM-AttrAux	0.6319	0.5902	0.6571	0.0677	0.0675	3.937	45.0	72.5
w/o AesAux	**0.6384**	**0.6025**	**0.6632**	0.0660	0.0666	3.747	75.5	84.5
w/o RelBias	0.6358	0.5975	0.6611	0.0664	0.0678	3.740	79.5	79.5
w/o Framing Conditioning	0.6307	0.5972	0.6519	0.0672	**0.0653**	**3.460**	73.0	66.5
NextShot	0.6373	0.5973	0.6630	**0.0659**	0.0670	4.068	—	—

Regional target prediction. NextShot improves regional geometry and frame prediction over NoChange, with box-overlap gains for both people and environmental regions (Table 1). On the shared centroid and visible-area fields, it also outperforms AR-8B; Qwen and Flash prefer its layouts in 79.5% and 81.0% of comparisons. Appendix C provides paired intervals and judgment counts.

Reference fit and layout preference. NextShot, w/o AesAux, and w/o RelBias achieve similar overall box IoUs, yet both evaluators prefer NextShot's layouts. These results support auxiliary aesthetic supervision and explicit relational biases for predicting preferred arrangements.

Framing conditioning improves box IoU and centroid accuracy, with the largest IoU gain in environmental regions. Both evaluators also prefer the resulting layouts, although the ablation achieves lower visible-area and frame errors.

5.3 Design Analysis

Aesthetic supervision. On the development set (Table 2), both evaluators favor auxiliary aesthetic supervision in the 4B predictor and the lightweight no-VLM model. The latter gain comes from attribute supervision without the rationale objective, extending the evidence beyond the full architecture.

Model configurations. Both evaluators prefer NextShot to NoVLM-Paired; their preferences diverge for w/o AesAux versus NoVLM-Paired under paired supervision alone. Appendix B reports backbone-scale and explicit-input analyses.

Alternative predictors. Both evaluators prefer NextShot to AR-8B and ZS-32B. On prepared inputs, NextShot has a median service time of 0.206 s versus 16.522 s for AR-8B (Appendix E).

Table 2: Development-set comparisons of auxiliary supervision, model configurations, and alternative predictors. Rates report A-over-B layout preference as a percentage of all planned judgments. Qwen evaluates 200 sources; Flash evaluates a 100-source subset.

Comparison (A vs. B)	A win rate (%)	
	Qwen	Flash
Auxiliary aesthetic supervision		
NextShot vs. w/o AesAux	75.5	86.0
NoVLM-AttrAux† vs. NoVLM-Paired	78.5	72.0
Model configurations		
w/o AesAux vs. NoVLM-Paired	43.0	61.0
NextShot vs. NoVLM-Paired	78.0	78.0
Alternative predictors		
NextShot vs. AR-8B	84.5	76.0
NextShot vs. ZS-32B	83.0	69.0

†NoVLM-AttrAux adds PortraitCraft attribute supervision without the rationale objective.

5.4 Guided Photography

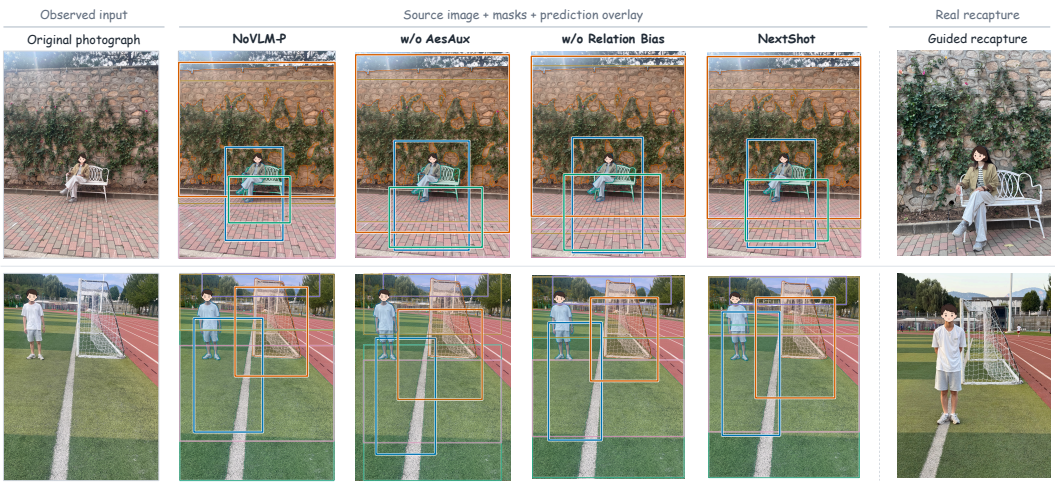

Figure 4: Predicted targets and guided recaptures on real inputs. Region identities and colors are consistent across predictions. The rightmost photographs were recaptured using NextShot targets. NoVLM-P and w/o Relation Bias denote NoVLM-Paired and w/o RelBias, respectively.

Targets on real inputs. Both evaluators prefer NextShot targets over NoVLM-Paired and w/o RelBias on the independently collected initial photographs, and both structured baselines over NoChange (Table 3(a)).

Photography protocol. Four photographers with everyday photography experience photographed two subjects. For each case, the photographer composed and captured the initial image normally before seeing guidance, without selecting it from multiple candidates. They then viewed only the NextShot target as regional boxes and semantics, adjusted the shooting setup, and took one guided recapture. All 100 pairs were retained and evaluated without additional photographic editing (Figure 4; Appendix D).

Five independent evaluators, uninvolved in photographing, posing, or image selection, each assessed all 100 anonymous A/B pairs, yielding 500 judgments. Pair-level majorities require at least three of five votes.

Table 3: Real-photography evaluation on 100 cases: (a) predicted targets, (b) human preferences from five evaluators per pair, and (c) model preferences for the photographs. Rates use all judgments or pairs.

(a) Predicted targets on real inputs

Comparison	Qwen win% ↑	Flash win% ↑
NextShot vs. NoVLM-Paired	**68.0**	**72.0**
NextShot vs. w/o RelBias	**67.0**	**86.0**
NoVLM-Paired vs. NoChange	**100.0**	**91.0**
w/o RelBias vs. NoChange	**97.0**	**94.0**

(b) Human evaluation of guided recaptures

Analysis level	N	Guided	Initial	No preference / no majority	Guided rate ↑
Individual judgments	500	**390**	36	74	**78.0%**
Pair-level $\geq 3/5$ majority	100	**89**	**0**	11	**89.0%**

(c) Model evaluation of guided recaptures

Photograph comparison	Qwen win% ↑	Flash win% ↑
Guided vs. initial ($N = 100$)	**83.0**	**90.0**

Photographic outcomes. Human evaluators prefer guided recaptures in 390 of 500 judgments (78.0%), with majorities in 89 of 100 pairs and no majority favoring an initial photograph (Table 3(b)). Both multimodal evaluators also favor the guided photographs (Table 3(c)). These results support the practical use of predicted regional targets as spatial guidance.

6 Discussion

Composition targets support human–model collaboration: the model proposes an inspectable spatial goal, and the photographer selects adjustments suited to the scene.

The current formulation describes two-dimensional visible geometry. Its physical attainability depends on scene depth, occlusion, camera field of view, and the available motion of the photographer and subject. Coupling the target with 3D reasoning or action planning is a natural extension. Training also depends on automatically extracted regions, partial correspondence, and weak person-framing labels; improved grounding and uncertainty-aware supervision could strengthen this foundation.

Extending guided recapture to broader scenes, more participants, and alternative guidance interfaces would clarify how these targets support photography across settings.

7 Conclusion

We introduced Region-Level Composition Target Prediction, which expresses a desired photographic arrangement through source-associated regional geometry and framing. AesRecon-CT and PortraitCraft-Aux make this task learnable from paired demonstrations and aesthetic annotations. NextShot combines relational encoding, multimodal context, and framing-conditioned decoding to predict the targets. Reference-based evaluation, layout comparisons, and a guided-recapture study demonstrate their value as structured predictions and spatial references for photographers. The framework connects understanding an observed scene to an explicit goal for the next photograph.

AI USE STATEMENT

AI tools were used to assist with brainstorming, coding and debugging, literature summarization, and language editing. All research decisions, experimental results, and conclusions were independently verified by the authors.

REPRODUCIBILITY STATEMENT

The core NextShot implementation, tests, input specifications, and anonymized data provenance records are available in an anonymous repository: https://anonymous.4open.science/r/artifact-4F8B/. Experiment configurations, data-construction scripts, and evaluation-subset identifiers will be released after the review process. Derived annotations will also be released where permitted by the upstream data licenses.

REFERENCES

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL Technical Report. *arXiv:2511.21631*, 2025. URL https://arxiv.org/abs/2511.21631.

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. In *CVPR*, pp. 15313–15322, 2026.

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 *ICLR*, 2026.

Tianxiang Du, Hulingxiao He, and Yuxin Peng. AesFormer: Transform Everyday Photos into Beautiful Memories. In *ICML*, 2026a.

Tianxiang Du, Hulingxiao He, and Yuxin Peng. Venus: Benchmarking and Empowering Multimodal Large Language Models for Aesthetic Guidance and Cropping. In *CVPR*, pp. 37766–37776, 2026b.

Johan Edstedt, David Nordström, Yushan Zhang, Georg Bökman, Jonathan Astermark, Viktor Larsson, Anders Heyden, Fredrik Kahl, Mårten Wadenbäck, and Michael Felsberg. RoMa v2: Harder Better Faster Denser Feature Matching. In *ECCV*, 2026.

Weixi Feng, Wanrong Zhu, Tsu-jui Fu, Varun Jampani, Arjun Akula, Xuehai He, Sugato Basu, Xin Eric Wang, and William Yang Wang. LayoutGPT: Compositional Visual Planning and Generation with Large Language Models. In *NeurIPS*, volume 36, pp. 18225–18250, 2023. doi: 10.52202/075280-0802.

Thorsten Hempel, Ahmed A. Abdelrahman, and Ayoub Al-Hamadi. 6D Rotation Representation for Unconstrained Head Pose Estimation. In *ICIP*, pp. 2496–2500, 2022. doi: 10.1109/ICIP46576.2022.9897219.

James Hong, Lu Yuan, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Learning Subject-Aware Cropping by Outpainting Professional Photos. In *AAAI*, volume 38, pp. 2175–2183, 2024. doi: 10.1609/aaai.v38i3.27990.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-Rank Adaptation of Large Language Models. In *ICLR*, 2022.

Yipo Huang, Leida Li, Pengfei Chen, Haoning Wu, Weisi Lin, and Guangming Shi. Multi-Modality Multi-Attribute Contrastive Pre-Training for Image Aesthetics Computing. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(2):1205–1218, 2025. doi: 10.1109/TPAMI.2024.3492259.

Tao Jiang, Xinchen Xie, and Yining Li. RTMW: Real-Time Multi-Person 2D and 3D Whole-body Pose Estimation. *arXiv:2407.08634*, 2024. URL https://arxiv.org/abs/2407.08634.

Zhaoyun Jiang, Jiaqi Guo, Shizhao Sun, Huayu Deng, Zhongkai Wu, Vuksan Mijovic, Zijiang James Yang, Jian-Guang Lou, and Dongmei Zhang. LayoutFormer++: Conditional Graphic Layout Generation via Constraint Serialization and Decoding Space Restriction. In *CVPR*, pp. 18403–18412, 2023.

Seung Hyun Lee, Jijun Jiang, Yiran Xu, Zhuofang Li, Junjie Ke, Yinxiao Li, Junfeng He, Steven Hickson, Katie Datsenko, Sangpil Kim, Ming-Hsuan Yang, Irfan Essa, and Feng Yang. Cropper: Vision-Language Model for Image Cropping through In-Context Learning. In *CVPR*, pp. 30010–30019, 2025.

Jiawan Li, Fei Zhou, Zhipeng Zhong, Jiongzhi Lin, and Guoping Qiu. Towards Smart Point-and-Shoot Photography. In *CVPR*, pp. 28242–28251, 2025.

Jiayu Li, Yixiao Fang, Tianyu Hu, Wei Cheng, Ping Huang, Zheheng Fan, Gang Yu, and Xingjun Ma. ShutterMuse: Capture-Time Photography Guidance with MLLMs. *arXiv:2606.25763*, 2026. URL https://arxiv.org/abs/2606.25763.

Jiawei Lin, Shizhao Sun, Danqing Huang, Ting Liu, Ji Li, and Jiang Bian. From Elements to Design: A Layered Approach for Automatic Graphic Design Composition. In *CVPR*, pp. 8128–8137, 2025.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. In *ECCV*, 2024.

Chancharik Mitra, Brandon Huang, Trevor Darrell, and Roei Herzig. Compositional Chain-of-Thought Prompting for Large Multimodal Models. In *CVPR*, pp. 14420–14431, 2024.

Jae Sung Park, Zixian Ma, Linjie Li, Chenhao Zheng, Cheng-Yu Hsieh, Ximing Lu, Khyathi Chandu, Quan Kong, Norimasa Kobori, Ali Farhadi, Yejin Choi, and Ranjay Krishna. Synthetic Visual Genome. In *CVPR*, pp. 9073–9086, 2025.

Sohan Patnaik, Rishabh Jain, Balaji Krishnamurthy, and Mausoom Sarkar. AesthetiQ: Enhancing Graphic Layout Design via Aesthetic-Aware Preference Alignment of Multi-modal Large Language Models. In *CVPR*, pp. 23701–23711, 2025.

Daiqing Qi, Handong Zhao, Jing Shi, Simon Jenni, Yifei Fan, Franck Dernoncourt, Scott Cohen, and Sheng Li. The Photographer's Eye: Teaching Multimodal Large Language Models to See, and Critique Like Photographers. In *CVPR*, pp. 24807–24816, 2025.

Yuyang Sha, Zijie Lou, Youyun Tang, Xiaochao Qu, Zheng Qu, Ben Xia, Haoxiang Li, Ting Liu, and Luoqi Liu. PortraitCraft: A Benchmark for Portrait Composition Understanding and Generation. *arXiv:2604.03611*, 2026. URL https://arxiv.org/abs/2604.03611.

Oriane Siméoni, Huy V. Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seung Eun Yi, Michaël Ramamonjisoa, Francisco Massa, Daniel Haziza, Luca Wehrstedt, Jianyuan Wang, Timothée Darcet, Théo Moutakanni, Leonel Sentana, Claire Roberts, Andrea Vedaldi, Jamie Tolan, John Brandt, Camille Couprie, Julien Mairal, Hervé Jégou, Patrick Labatut, and Piotr Bojanowski. DINOv3. *Transactions on Machine Learning Research*, 2026.

Yukun Su, Yiwen Cao, Jingliang Deng, Fengyun Rao, and Qingyao Wu. Spatial-Semantic Collaborative Cropping for User Generated Content. In *AAAI*, volume 38, pp. 4988–4997, 2024. doi: 10.1609/aaai.v38i5.28303.

Sheyang Tang, Armin Shafiee Sarvestani, Jialu Xu, Xiaoyu Xu, and Zhou Wang. Aesthetic Camera Viewpoint Suggestion with 3D Aesthetic Field. In *CVPR*, pp. 8278–8287, 2026.

Junsheng Wang, Nieqing Cao, Yan Ding, Mengying Xie, Fuqiang Gu, and Chao Chen. SKE-Layout: Spatial Knowledge Enhanced Layout Generation with LLMs. In *CVPR*, pp. 19414–19423, 2025.

Haoning Wu, Zicheng Zhang, Weixia Zhang, Chaofeng Chen, Liang Liao, Chunyi Li, Yixuan Gao, Annan Wang, Erli Zhang, Wenxiu Sun, Qiong Yan, Xiongkuo Min, Guangtao Zhai, and Weisi Lin. Q-Align: Teaching LMMs for Visual Scoring via Discrete Text-Defined Levels. In *ICML*, volume 235, pp. 54015–54029, 2024.

Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth Anything V2. In *NeurIPS*, 2024.

Zhiwei Yang, Yuanchen Wu, Nan Zhang, Yucong Meng, Ke Yan, and Shouhong Ding. Scene Graph Thinking: Reinforcing Structured Visual Reasoning for Multimodal Large Language Models. In *ICML*, 2026.

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*, volume 38, 2025. doi: 10.52202/085713-3020.

Zhiyuan You, Ke Wang, He Zhang, Xin Cai, Jinjin Gu, Tianfan Xue, Chao Dong, and Zhoutong Zhang. PhotoFramer: Multi-modal Image Composition Instruction. In *CVPR*, pp. 10197–10207, 2026.

Ke Zhang, Tianyu Ding, Jiachen Jiang, Tianyi Chen, Ilya Zharkov, Vishal M. Patel, and Luming Liang. ProCrop: Learning Aesthetic Image Cropping from Professional Compositions. In *AAAI*, volume 40, pp. 12600–12608, 2026. doi: 10.1609/aaai.v40i15.38255.

Zhaoran Zhao, Peng Lu, Anran Zhang, Peipei Li, Xia Li, Xuannan Liu, Yang Hu, Shiyi Chen, Liwei Wang, and Wenhao Guo. Can Machines Understand Composition? Dataset and Benchmark for Photographic Image Composition Embedding and Understanding. In *CVPR*, pp. 14411–14421, 2025.

A Data, Supervision, and Training Details

This appendix expands the construction of AesRecon-CT and PortraitCraft-Aux in Section 3, and documents the training and evaluation sets used in Section 5.1.

A.1 Photographic data preparation

Two reviewers screen AesRecon source–reference groups for portrait relevance and usable photographic content. The screening removes problematic composites and screenshots. Visible borders or opaque camera-interface bars are conservatively cropped when they alter the effective image canvas; dependent regional geometry is updated to that canvas. Each source photograph and all its references remain one group throughout preparation and partitioning.

The reference photographs provide demonstrations of alternative compositions of the same photographic event. We preserve their regional arrangements rather than creating a target by averaging references. Region extraction, correspondence, and weak person-framing labels are produced by the automatic pipeline after image-level screening. Thus, AesRecon-CT adds source-associated regional targets to the original photographic demonstrations.

A.2 Evaluation membership and partitioning

Table 4 summarizes the sets used in the paper. The 450-source development set contains 647 valid references and is used for checkpoint selection and design comparisons. Its selection emphasizes visible composition changes: 446 groups meet person-position or person-scale criteria, and four meet an environmental-change criterion. Source groups assigned to development, together with identified linked groups, are removed from paired training.

Table 4: Data used for training and evaluation. Counts denote source photographs unless otherwise stated. Randomly sampled layout-preference subsets are held fixed across method comparisons.

Set	Size	Supervision or evaluation unit
AesRecon-CT training	2,915	4,268 valid photographic references
PortraitCraft-Aux training	25,164	Images with attribute scores and rationales
Development	450	647 valid references; checkpoint selection
Test reference agreement	781	One eligible reference per source
Test layout preference	200	Random sample shared by Qwen and Flash
Development layout preference	200 / 100	Random Qwen sample / random Flash subset
Guided photography	100	Initial–recaptured pairs; 500 human judgments

Of 792 groups in the original AesRecon test assignment, 791 have valid source inputs and 781 have eligible reference geometry. The remaining ten valid inputs have no eligible geometric reference and are excluded from reference agreement. The 781 scored sources belong to 780 known near-duplicate components. Their person and environment metrics use 740 and 727 sources containing valid regions of the respective type.

Test-set layout comparisons use a fixed random sample of 200 sources, drawn with seed 20260922 from the 790 known components of the 791 valid inputs, one source per sampled component. The framing-conditioning comparison uses this same subset. Development layout evaluation randomly samples 200 sources from the development set for Qwen, with a random 100-source subset for Flash. These samples are held fixed across comparisons and share no source IDs with the test preference subset. The photography study is collected independently of AesRecon-CT.

Source groups remain intact across partitions. File, pixel, and perceptual-similarity audits find no known overlap between test sources and current paired-training, auxiliary-training, or development data. Earlier models had been run on the retained test assignment; the current checkpoints and evaluation protocol were fixed before scoring.

A.3 REGIONAL EVIDENCE AND CORRESPONDENCE

Source evidence. Qwen3-VL proposes composition-relevant people and environmental components. Grounding DINO localizes each phrase, with its Qwen box as a fallback; SAM 3 produces a box-prompted mask. The mask supplies the enclosing box, visible centroid (the mean position of mask pixels), and visible-area fraction (mask pixels divided by image area). Coordinates are normalized by the corresponding canvas. Source and reference images are processed independently. Table 5 lists the frozen perception models.

Table 5: Perception models used to construct regional evidence.

Evidence	Model / configuration
Region proposals	Qwen3-VL-32B-Instruct (Bai et al., 2025)
Phrase grounding	Grounding DINO Base (Liu et al., 2024)
Visible masks	SAM 3 (Carion et al., 2026), `sam3.pt` image checkpoint
Regional appearance	DINOv3 ViT-L/16 (Siméoni et al., 2026) from the official RoMaV2 checkpoint (Edstedt et al., 2026)
Relative depth	Depth Anything V2 (Yang et al., 2024), Base (ViT-B), input size 518
Body keypoints	RTMW-L (Jiang et al., 2024), Cocktail14, 256×192, WholeBody133
Head orientation	6DRepNet (Hempel et al., 2022), 300W-LP/AFLW2000 checkpoint

DINOv3 features concatenate zero-indexed layers 17 and 23 into 2,048 dimensions. Region and context crops use margins of 5% and 20% per side, respectively, and a 448-pixel long edge. Region pooling uses the mask; context pooling uses the valid crop. Each layer's vector and the concatenation are L_2-normalized. Body evidence packs 133 keypoints with normalized coordinates, confidence, and mask support, plus one availability flag (533 dimensions). Depth statistics and body/head orientation enter source representations.

Source–reference association. Let u_{ij} and v_{ij} be region and context cosine similarities clipped to $[0, 1]$, z_{ij} semantic compatibility, and p_{ij} indicate that both regions are marked as the primary person. Matching uses

$$S_{ij} = \min(1,\ 0.75u_{ij} + 0.15v_{ij} + 0.10z_{ij} + 0.035p_{ij})\,.$$

Semantic evidence combines group agreement, content-tag Jaccard overlap, and description-token Jaccard similarity with weights $0.35/0.45/0.20$; person–person pairs use $0.75 + 0.25J$, where J is token Jaccard similarity. Missing content tags receive overlap 0.3. Content tags follow a fixed lexical taxonomy. Person/nonperson and incompatible content types are excluded; person-group or body-part granularity conflicts remain unconfirmed. Position, scale, shape, and depth similarity do not enter the score.

Compatible pairs become candidates if $u_{ij} \geq 0.40$ and $S_{ij} \geq 0.48$, or if $p_{ij} = 1$ or $z_{ij} \geq 0.78$. Partial bipartite assignment maximizes total utility $S_{ij} - 0.60$, with dummy unmatched options of zero utility. An assigned pair is accepted only if its region similarity reaches 0.70 (0.65 for primary-person pairs; 0.72 for continuous-environment content), its score reaches 0.72 (0.695 for primary-person pairs), and forbidding that edge decreases optimal assignment utility by at least 0.045. For unmatched endpoints, up to three candidate edges are retained; a competing retained edge sharing an endpoint and lying within 0.045 of an accepted score makes that correspondence ambiguous.

Accepted reference regions supply geometry in their own frame to the associated source identities. Unmatched or ambiguous source regions remain context and prediction variables, with their geometric supervision masked; reference-only regions add no output identity.

Complete targets and weak framing. Each reference retains its own validity mask. Candidate references are compared on their common valid region set, and supervision uses all valid regions of the selected reference (Section 4.4). Reference-body evidence yields weak framing candidates: `full_body`→full and `waist_up`/`chest_up`→half; other candidates, including `knee_up`, `head_closeup`, and `unknown`, are unsupervised. Labels apply only to valid matched person regions.

A.4 AUXILIARY AESTHETIC ANNOTATIONS

PortraitCraft-Aux retains seven composition-related attributes: compositional principles, depth and layering, visual-center stability, visual flow, structural support, negative space, and subject completeness. Scores on $[0, 10]$ are divided by 10 for regression; unavailable scores are masked. The attribute loss averages smooth-L_1 errors ($\beta = 0.1$) over available attributes within each image. One available attribute rationale per image supplies answer-token cross-entropy, with at most 128 answer tokens. Attribute readouts precede the rationale in the causal sequence. These annotations describe the observed photograph; paired references supply the desired geometry.

A.5 TRAINING AND CONTROLLED VARIANTS

Objective scales. Let H denote smooth-L_1 with $\beta = 1$. The per-region error for reference k is

$$\ell_i^{(k)} = \frac{1}{4}\sum_{d=1}^{4} H\left(\frac{\hat{b}_{id} - b_{id}^{(k)}}{0.1}\right) + \frac{1}{2}\sum_{d=1}^{2} H\left(\frac{\hat{c}_{id} - c_{id}^{(k)}}{0.1}\right) + H\left(\frac{\log \hat{a}_i - \log a_i^{(k)}}{0.7}\right).$$

The frame error is $D_\rho = H((\hat{\rho} - \rho^{(k)})/0.3)$. Area values are lower-bounded by 10^{-8} before taking logarithms. These normalized errors are used in both reference selection and supervision, with the validity sets and role averaging defined in Section 4.4. The weights are $\lambda_\rho = 1$, $\lambda_f = 0.1$, $\lambda_{\text{attr}} = 0.25$, and $\lambda_{\text{rat}} = 0.05$; the coarse-framing term is cross-entropy over available labels.

Architecture and optimization. Table 6 specifies the main 4B architecture and training settings, with d_{VLM} denoting the backbone hidden size.

Table 6: Architecture and optimization settings for NextShot.

Setting	Value
Input dimensions	Appearance 2,048; geometry 14; orientation 12; body 533
Relational hidden size / heads	256 / 4
Relational layers	1 source encoder; 2 initial decoder + 1 conditioned decoder
Shared projector P	LayerNorm $\to$ Linear $\to$ GELU $\to$ Linear; $256 \to d_{\text{VLM}} \to d_{\text{VLM}}$
LoRA	Language-layer query/value projections; rank 8, $\alpha = 16$
Optimizer	AdamW; weight decay 0.01
Learning rates	Structured modules: 2×10^{-4}; LoRA: 2×10^{-5}
Schedule	5% warmup, then cosine decay to 10% of initial rates
Effective batch per update	4 paired sources; 5 or 6 auxiliary images
Image budget / seed	65,536–262,144 input pixels; seed 20260916

Mixing supervision. A shuffled single pass over the 25,164 auxiliary images is distributed evenly across the planned paired-data updates. Each update accumulates paired and weighted auxiliary gradients before one optimizer step. Each task is normalized by its own source count; rationale loss averages over images with available rationales. Training uses two GPUs, with microbatches of two paired sources and up to three auxiliary images per GPU.

Training budgets and variants. Structured 4B/8B runs allow up to six passes over the paired data; the two no-VLM configurations allow up to 20. Each configuration selects its own checkpoint using the development set. NextShot selects epoch 4, w/o AesAux and w/o RelBias select epoch 6, NextShot-8B selects epoch 5, and NoVLM-Paired/NoVLM-AttrAux select epochs 3/4. The selected NextShot checkpoint has 2,916 updates and 16,776 auxiliary-image visits.

The w/o RelBias variant removes the explicit bias modules in the source encoder and three decoder layers, a total of 1,936 parameters, while retaining ordinary attention and region interaction. The w/o Framing Conditioning variant removes predicted frame and person-framing inputs to geometry decoding while retaining their heads, losses, and relational layers. It uses the same training-budget ceiling and early-stopping/selection rules as NextShot, stops at epoch 5, and selects epoch 3 (2,187 updates).

B Development-Set Design Comparisons

B.1 Auxiliary supervision

Table 2 in the main text shows that auxiliary aesthetic supervision improves layout preference in both the 4B architecture and the lightweight no-VLM predictor. For NoVLM-AttrAux versus NoVLM-Paired, Flash records 72 wins, 26 losses, and two ties over 100 sources, giving a 72.0% preference rate; Qwen reports 78.5% over 200 sources.

B.2 Backbone scale

NextShot-8B replaces the 4B backbone with an 8B backbone under the same joint supervision objective. Table 7 reports the development-set comparison, which yields differing preferences across the two evaluators.

Table 7: Development-set backbone-scale comparison. Rates report preference for the 4B NextShot model over NextShot-8B as a percentage of all planned judgments. Qwen evaluates 200 sources; Flash evaluates the same 100-source subset used in Table 2.

Comparison (A vs. B)	Qwen A-win%	Flash A-win%
NextShot vs. NextShot-8B	71.0	37.0

B.3 Contribution of explicit source evidence

We additionally examine four independently trained variants that remove explicit source-evidence channels before text and graph construction. The targets and cached assets are unchanged. Depth removal masks depth statistics as unavailable; orientation removal zeros the explicit orientation array. Body-evidence removal zeros the 533-dimensional body input and removes its textual availability cues, while retaining the person-framing targets and prediction heads. Semantic removal clears region descriptions and semantic compatibility, uses generic region IDs, and retains the person-routing flag needed by the decoder. Image inputs and regional appearance features remain available.

Table 8: Development-set preference for NextShot over explicit-input variants. W/L/T/F denote NextShot wins, losses, ties, and failures. Rates use all planned judgments: 200 for Qwen and 100 for Flash.

Input removed	Qwen W/L/T/F	Win%	Flash W/L/T/F	Win%
Depth	149/51/0/0	74.5	86/11/2/1	86.0
Orientation	152/47/1/0	76.0	81/15/3/1	81.0
Body evidence	104/96/0/0	52.0	67/29/4/0	67.0
Semantics	127/73/0/0	63.5	72/20/6/2	72.0

Depth, orientation, and semantic evidence yield consistent preference gains across both evaluators (Table 8). The body-evidence comparison is weaker under Qwen. These results complement the main structural ablations by showing how source evidence contributes to the predicted arrangement. They concern explicit input channels; the photograph can still contain related visual cues.

C Additional Test-Set Evidence

C.1 Geometric agreement and aggregation

Regional metrics are averaged over valid regions within a source, then equally across sources. Box IoU measures enclosing-box overlap; centroid MAE averages absolute error over the two normalized visible-centroid coordinates; visible-area MAE is the absolute error in mask area as a fraction of the target frame. Frame relative error is $100|\hat{r}/r - 1|$, where r is the reference width-to-height ratio. Person/environment results average only sources with valid regions of the corresponding type.

For multi-reference development examples, a single complete reference is selected on common valid regions before scoring all valid regions in that reference. Each of the 781 test sources has one eligible reference, so reference selection does not change its target. AR-8B predicts the shared centroid and visible-area fields. One failed AR-8B output in the scored set uses the frozen NoChange fallback; structured predictions require no fallback.

Table 9: Paired test-set differences between NextShot and two prediction baselines on 781 sources. Differences are NextShot minus the baseline. Positive IoU and negative error differences favor NextShot. Frame differences are percentage points (pp).

Baseline	Metric	Difference	95% paired interval
NoChange	Box IoU	$+0.073484$	$[+0.065551, +0.081425]$
	Centroid MAE	-0.008597	$[-0.010243, -0.006985]$
	Visible-area MAE	-0.011043	$[-0.014265, -0.007914]$
	Frame RE (pp)	-5.780025	$[-6.613363, -4.981581]$
AR-8B	Centroid MAE	-0.002070	$[-0.003463, -0.000657]$
	Visible-area MAE	-0.005666	$[-0.008519, -0.002880]$

Table 9 supports learned changes in both regional position and visible extent. Excluding the single AR-8B fallback yields centroid and area differences of -0.002088 and -0.005681, respectively, preserving the direction of the comparison. Paired intervals use 10,000 percentile bootstrap resamples of the 780 known source components, preserving source weights. They describe sample uncertainty for the fixed checkpoints and are pointwise intervals.

C.2 Complete-layout preference

Table 10 expands the preference columns of Table 1. Each method comparison uses the same 200-source subset. Wins are counted against all 200 planned judgments, retaining ties and failures in the denominator. The full predictor is favored over the source layout, AR-8B, and each of the three ablations by both evaluators. NoVLM-AttrAux yields different preferences across evaluators.

Table 10: Test-set preference for NextShot over each row method. W/L/T/F are wins, losses, ties, and failures. Each evaluator has 200 planned judgments per comparison.

Compared method	Qwen W/L/T/F	Win%	Flash W/L/T/F	Win%
NoChange	179/21/0/0	89.5	154/42/1/3	77.0
AR-8B	159/41/0/0	79.5	162/33/3/2	81.0
NoVLM-AttrAux	90/108/2/0	45.0	145/35/16/4	72.5
w/o AesAux	151/47/2/0	75.5	169/20/9/2	84.5
w/o RelBias	159/38/3/0	79.5	159/21/17/3	79.5
w/o Framing Conditioning	146/48/6/0	73.0	133/38/27/2	66.5

C.3 FRAMING-CONDITIONED REGIONAL PREDICTION

Conditioning regional geometry on the predicted frame and person framing improves overall box agreement and centroid accuracy (Table 11). The largest IoU gain is in environmental regions, whose arrangement depends on the canvas and its relationship to the person. Person IoU changes little. The ablation has lower frame and visible-area errors, while both layout evaluators favor the complete model's arrangements.

Table 11: Framing-conditioning ablation on 781 test sources. Differences are NextShot minus the ablation, with pointwise 95% intervals from the paired component bootstrap. Frame differences are in percentage points (pp).

Metric	Difference	95% paired interval
Box IoU	$+0.006652$	$[+0.003781, +0.009460]$
Box-coordinate MAE	-0.001022	$[-0.001564, -0.000473]$
Centroid MAE	-0.001291	$[-0.002048, -0.000530]$
Visible-area MAE	$+0.001714$	$[+0.000011, +0.003372]$
Frame RE (pp)	$+0.6083$	$[+0.4512, +0.7721]$
Person IoU	$+0.000080$	$[-0.003338, +0.003343]$
Environment IoU	$+0.011134$	$[+0.007282, +0.015045]$
Person/environment macro IoU	$+0.005607$	$[+0.003058, +0.008148]$

For this comparison, Qwen records 146 wins, 48 losses, and six ties for NextShot. Flash records 133 wins, 38 losses, 27 ties, and two refusals. Each source contributes one final judgment per evaluator. The two Flash refusals remain in the 200-source denominator. Under Qwen, NextShot wins 75 of the 100 cases displayed in position A and 71 of the 100 displayed in position B. These use different images in the two positions.

D GUIDED PHOTOGRAPHY: PROTOCOL AND ADDITIONAL RESULTS

D.1 FROM SOURCE REGIONS TO A GUIDED RECAPTURE

The study contains 100 initial photographs with 565 source regions. NextShot, NoVLM-Paired, and w/o RelBias each produce layouts for all 100 inputs with aligned source identities. The layout comparisons in Table 3(a) use these same images. Four photographers with everyday photography experience photograph two subjects. For each case, they first compose and capture an image without seeing guidance, then view the NextShot target boxes and regional semantics and take one guided recapture. All 100 pairs are retained, with no additional photographic editing before evaluation.

D.2 HUMAN PREFERENCES ACROSS EVALUATORS

Five evaluators, separate from the photographers, subjects, and image selection, independently assess all 100 anonymous A/B pairs. Pair order is randomized, and each evaluator sees the guided photograph in A and B for 50 pairs each. The question concerns overall composition preference, with options for either photograph, similar quality, or inability to judge. Similar-quality and unable-to-judge responses are pooled as no preference; pair-level no majority means neither photograph receives at least three votes.

Table 12: Human preferences by evaluator, with 100 photograph pairs per row. No preference combines ties and unable-to-judge responses. Evaluator identities are anonymized.

Evaluator	Guided	Initial	No preference	Guided rate
R1	89	11	0	89.0%
R2	58	7	35	58.0%
R3	78	4	18	78.0%
R4	82	4	14	82.0%
R5	83	10	7	83.0%
All judgments	390	36	74	78.0%

Every evaluator favors the guided photographs in more than half of their judgments (Table 12). Of the 100 pairs, 33 receive five guided votes, 38 receive four, 18 receive three, eight receive two, and three receive one. Thus, 71 pairs receive at least four guided votes, and 89 receive a guided majority. Eleven pairs have no option supported by at least three evaluators; no pair has an initial-photo majority. The 500 judgments refer to 100 photograph pairs, with repeated participants and locations; the reported proportions summarize this study.

D.3 AUTOMATIC PHOTO COMPARISONS

Table 3(c) reports Qwen and Flash preferences for the actual initial and recaptured photographs, using the photo prompt in Appendix F.4. Each sees one anonymous presentation per pair, with the guided photograph in A/B equally often. These comparisons use only the two photographs and are separate from the regional-layout comparisons.

Table 13: Automatic preference on the same 100 photograph pairs. Rates use all planned pairs, including invalid responses. The Qwen row uses the original strict parsing rule.

Evaluator	Guided	Initial	Tie	Invalid	Guided rate
Qwen	83	2	0	15	83.0%
Flash	90	10	0	0	90.0%

Both evaluators favor guided recaptures (Table 13), complementing the human preferences. Qwen's 15 invalid responses exceed the prompt's 25-word reason limit and remain in the denominator. For Flash, the guided photograph wins 45 of 50 cases in each display position.

E MEASURED RESOURCE USE

E.1 PREDICTION-SERVICE LATENCY

We measure the implemented predictors on the same 50 prepared development inputs using one A800 GPU, batch size one, a fixed input order, and three separate warm-up inputs. Each measured input is processed once. Service timing includes source-package loading and validation, RGB decoding, preprocessing, device transfer, forward prediction or generation, and conversion or parsing of the numerical output. Upstream region perception and feature extraction, model loading, and guidance rendering are outside this timing interval.

Table 14: Measured resource use. (a) Prediction service on the same 50 prepared inputs; every listed system produces 50/50 valid outputs. Peak memory is PyTorch allocated memory. (b) Actual training-loop resource use; memory is the peak per GPU.

(a) Prediction service: one A800, batch size one

Method	Median (s)	P95 (s)	Forward / generation (s)	Peak (GiB)
NoVLM-Paired	0.087	0.124	0.009	0.051
NoVLM-AttrAux	0.059	0.080	0.009	0.051
w/o AesAux	0.206	0.306	0.144	8.729
NextShot	0.206	0.268	0.144	8.729
NextShot-8B	0.235	0.310	0.167	15.692
AR-8B	16.522	26.073	16.455	19.506

(b) Training-loop measurements

Method	Trainable (M)	GPUs	Peak (GiB)	GPU-hours
NoVLM-Paired	4.833	1	0.43	0.028
NoVLM-AttrAux	4.835	1	2.40	0.042
w/o AesAux	15.924	2	27.99	1.886
NextShot	15.949	2	43.71	4.660
NextShot-8B	28.265	2	43.03	6.049
AR-8B	3.834	2	54.56	19.225

NextShot's median prediction-service time is 0.206 s, compared with 16.522 s for AR-8B (Table 14(a)), about an 80-fold difference for the measured systems. Its allocated inference-memory peak is 8.729 GiB. The structured decoder produces the regional targets in one forward pass; AR-8B generates the shared centroid and area fields autoregressively. The measurements retain each system's implemented output format and generation settings.

E.2 TRAINING RESOURCES

Table 14(b) reports actual executed training loops, including loop-internal validation and checkpoint I/O. NextShot has 15.949 million trainable parameters within a 4.454-billion-parameter training model and uses 4.660 GPU-hours. Counts exclude upstream perception modules; training time excludes base-model pretraining and data preparation. The no-VLM runs use a 20-epoch ceiling, the structured VLM runs use a six-epoch ceiling, and AR-8B follows its own update schedule. Reported costs cover executed loops rather than stopping the accounting at the selected checkpoint.

F AUTOMATIC EVALUATION PROTOCOL AND PROMPTS

F.1 INPUTS, PRESENTATION, AND AGGREGATION

The evaluators are Qwen3.8-27B (Qwen) and Gemini-3.8-Flash (Flash). Layout requests provide the source photograph, common numbered-region image, candidate A, and candidate B, followed by region identities and geometric facts. Coordinates are normalized within each candidate's frame. Shared identities contain region IDs, semantic groups, and neutral descriptions; layout input labels are Chinese. No reference photograph or method identity is supplied.

For test layouts, NextShot occupies A/B for 100 sources each under Qwen and A throughout under Flash. Real-input comparisons use a 50/50 assignment under Qwen and a fixed A assignment for the left-hand method under Flash. Development runs use fixed candidate orders within each comparison. Each pair receives one presentation per evaluator; position groups contain different images.

The four criteria output A/B/tie/uncertain. The candidate with more directional votes wins; equal A/B counts yield a tie. Criterion-level tie and uncertain responses do not vote. Overall expression contributes one vote. Invalid responses are recorded separately, and preference rates divide wins by all planned judgments. The four decisions form one pairwise judgment per source.

F.2 COMPLETE-LAYOUT SYSTEM PROMPT

The following English system prompt is shared by the two evaluators. The source photograph supplies scene content, while the layouts describe candidate geometric targets. Test body-goal labels are `not_specified`; ZS-32B development comparisons restrict goals to the same primary person.

> Compare the photographic composition of anonymous layouts A and B for the same scene. Use the scene photograph and numbered regions to understand the content, and the candidate layouts and geometric facts to assess each arrangement. Identify the main subject from the scene rather than its region number. Apply the same standards to both candidates, regardless of presentation order or resemblance to the source photograph. Use your knowledge of photographic composition to compare the two images, grounding each judgment in visible features rather than mechanically applying rules.
>
> Judge these four criteria independently:
>
> 1. subject_hierarchy: How clearly does the arrangement establish a focal point and distinguish primary from supporting elements? Consider relative scale, placement, and competing attention. Assess how subject scale supports the scene and preserves useful context.
>
> 2. spatial_balance: How well are visual weights distributed across the frame? Consider subject placement, supporting elements, directional flow, and the frame's proportions. Both symmetry and purposeful asymmetry can be effective; balance is not equal area or mandatory centering.
>
> 3. subject_environment: How well does the subject relate to its setting? Consider relative scale, placement, separation, and whether the surroundings support the scene's meaning without competing unnecessarily with the subject.
>
> 4. overall_expression: How coherent and effective is the composition as a whole? Consider clarity, visual flow, and the interaction of the elements. Make a whole-composition judgment rather than counting the other three votes.
>
> The layouts are geometric proposals, not finished photographs. Boxes show extents, crosses show centroids, and the supplied visible-area values describe region size. Use each candidate's own frame. Treat body-framing labels as proposed goals; not_specified means no stated goal. Judge compositional effects supported by these materials; exact target contours, poses, and occlusion are not shown.
>
> For each criterion, choose A or B when one arrangement is better supported, tie when they are similarly effective, or uncertain when missing information prevents that particular judgment. Ordinary trade-offs can be weighed without requiring uncertain.
>
> Return only one JSON object with exactly four keys: subject_hierarchy, spatial_balance, subject_environment, overall_expression. Each value must be A, B, tie, or uncertain as a JSON string. Output no explanation or additional fields.

F.3 CENTROID/AREA PROMPT VARIANTS

AR-8B comparisons replace only the material-description paragraph of the preceding prompt. The criteria, decision options, and output requirements remain unchanged. Both candidates retain their own frame: NextShot uses its prediction and AR-8B uses the source frame.

Qwen replacement.

> The layouts are geometric proposals, not finished photographs. Crosses show centroids; circles are scaled symbols encoding the supplied visible-area values, not object outlines or bounding boxes. Use each candidate's own frame. Both candidates provide only region centers, visible-area fractions, and frame proportions. No target boxes or body-framing goals are supplied; do not infer them from the symbols. Judge compositional effects supported by these materials; exact target contours, poses, and occlusion are not shown.

Flash replacement.

> The layouts are geometric proposals, not finished photographs. Crosses show centroids. Circles are scaled area markers: each circle occupies one sixteenth of the supplied visible-area fraction. Use the numerical area values for region size and each candidate's own frame. Circle outlines are not object extents; their overlap or clipping does not establish occlusion or object truncation. Only center, visible area, and frame are provided; target boxes and body-framing goals are not available. Judge compositional effects supported by these materials.

F.4 REAL-PHOTOGRAPH SYSTEM PROMPT

The photo comparisons in Appendix D provide the following prompt and two anonymous image inputs, Photo A and Photo B.

> Compare the photographic composition of Photo A and Photo B fairly, regardless of presentation order. Use your knowledge of photographic composition to compare the two images, grounding each judgment in visible features rather than mechanically applying rules.
>
> Assess four criteria independently:
>
> - focus: Clarity of the main subject and hierarchy of attention, considering placement, scale, separation, and competing elements.
> - balance: Distribution of visual weight and space, including supporting elements and negative space. Both symmetry and purposeful asymmetry can work; centering is not required.
> - context: How viewpoint, background, relative scale, and depth connect the subject to meaningful surroundings.
> - coherence: Overall effectiveness of framing, visual flow, and the arrangement of elements. Assess the whole image rather than counting the other three judgments.
>
> Consider pose, gaze, light, and color through their effects on composition. Greater brightness, sharpness, saturation, or a more appealing expression alone does not establish better composition.
>
> For each criterion, choose "A" or "B" for the better composition, "tie" for similarly effective compositions, or "uncertain" when image evidence is insufficient. Weigh ordinary trade-offs.
>
> Return only one JSON object with keys "focus", "balance", "context", "coherence", and "reason". The first four values must be "A", "B", "tie", or "uncertain". Give one English sentence of at most 25 words in "reason", identifying the decisive visible difference or obstacle to comparison.

The four categorical decisions use the same directional-vote aggregation. The reason does not vote; the strict parser enforces its 1–25-word limit using whitespace-delimited words. Invalid records remain in the planned denominator.