Drag, Infer, Reproject: Grounding LLMs through Spatial Interaction for Image Clustering

Yang Liu* Virginia Tech
Xuxin Tang† Virginia Tech
Jiahao Xu‡ Virginia Tech
Chris North§ Virginia Tech

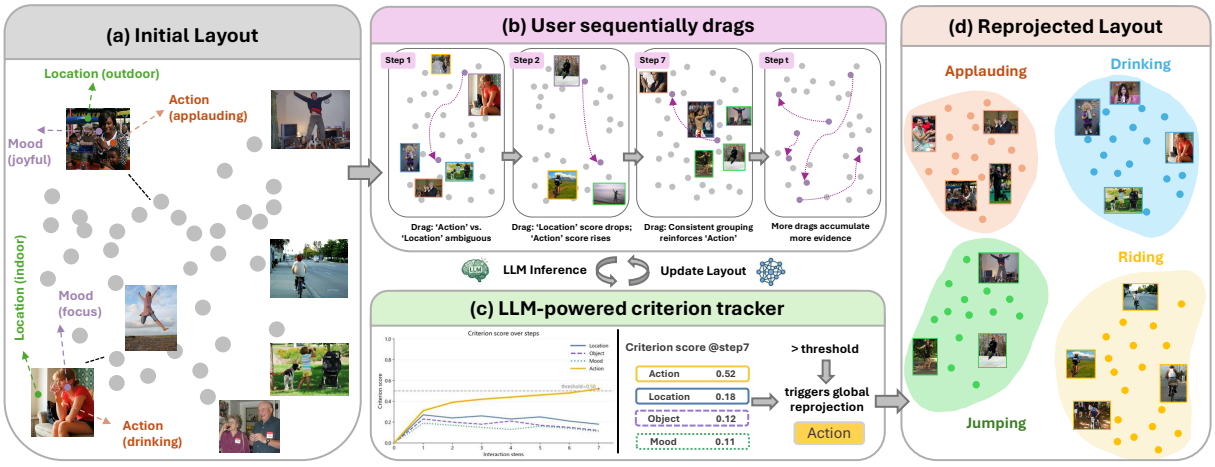

Figure 1: CriterionSI infers and refines the user's clustering criterion from incremental drag interactions. (a) An initial EVA-CLIP projection shows images in a semantically mixed layout. (b) The user drags images; the tracker considers multiple candidate criteria. (c) Once consistent evidence accumulates, the tracker commits to a criterion and assigns values to all images via LLMs. (d) Criterion-guided reprojection reshapes the entire layout, placing images into semantically coherent clusters.

Abstract

Dimension reduction and semantic interaction support image clustering by making similarity structure visible and manipulable. Existing semantic interaction methods encode users' clustering criterion (a user-interpretable semantic dimension, e.g., *action*, *location*, or *mood*) from direct manipulation to steer reprojection, giving users direct control over the resulting layout. Yet they typically depend on learned embeddings or a predefined criterion. In practice, users' clustering criterion often emerges gradually and becomes refined through interaction rather than being fully clear at the outset. In this work, we present **CriterionSI** (**Criterion**-guided **S**emantic **I**nteraction), a method that translates incremental drag interactions into criterion-guided reprojection. CriterionSI uses large language models to infer and refine the clustering criterion from sequential user drags, while grounding semantic interpretation in human-provided feedback rather than fixed prior assumptions. CriterionSI combines the inferred criterion with local drags to guide global reprojection. The simulation-based evaluation and usage scenario demonstrate that CriterionSI can discover and refine the target criterion from sequential interactions and progressively produce criterion-aligned clustering layouts. Our code and data are available at: https://github.com/4C79/CriterionSI.

Index Terms: Semantic Interaction, Dimensionality Reduction, Criterion Inference, Image Clustering, Large Language Models.

*e-mail: yangliu07@vt.edu
†e-mail: xuxintang@vt.edu
‡e-mail: jiahao@vt.edu
§e-mail: north@vt.edu

1 Introduction

Visual analytics enables users to gain insight into complex data through interactive visual exploration [7, 4]. In this context, image analysis often involves clustering collections of images to discover similarity relationships and identify meaningful structure in high-dimensional data [6, 8]. Dimension reduction (DR) supports this process by mapping high-dimensional embeddings into 2D spaces, where spatial proximity serves as a perceptual cue for grouping and pattern discovery [19, 11, 5].

In practice, users rarely begin with a fully specified clustering criterion [14, 13]. Instead, they express partial and evolving intent through sequential interactions with the projection [15, 3, 22]. Each individual manipulation, such as dragging an image toward a group, may admit multiple plausible semantic interpretations, making single-step reasoning inherently ambiguous [21]. As a result, the clustering criterion becomes identifiable through the accumulation and refinement of evidence across interaction steps.

Recent advances in vision-language embeddings and Large Language Models (LLMs) provide rich visual-semantic embeddings and strong capabilities in semantic reasoning, making them effective for interpreting and organizing high-level features [25]. Building on these capabilities, recent work has explored using LLMs to steer data clustering and visual representations, for example, by incorporating textual criteria or guiding model behavior through prompts and interaction [9, 12, 26]. In parallel, deep learning-based methods extend semantic interaction (SI) to neural models [2, 1], including image projections that learn updated embedding weights from user demonstrations [10, 20], though the organizing criterion remains implicit in the learned representation. This gap motivates making the clustering criterion explicit and reasoning about it with LLMs while preserving direct spatial interaction.

In this work, we present CriterionSI, a method for criterion-guided image clustering from sequential user interactions, as shown

in Fig. 1. Here, a clustering criterion refers to any semantic dimension that organizes images into groups. CriterionSI treats each drag as partial evidence of an evolving clustering criterion and uses an LLM-powered criterion tracker to infer, verify, and stabilize candidate criteria over interaction steps. Once the criterion stabilizes, CriterionSI assigns semantic values to all images with LLMs and combines the resulting criterion constraints with visual similarity and drag constraints to guide global reprojection, aligning the layout with the user's clustering intent. Our contributions are: (1) An LLM-powered criterion tracker that infers clustering criterion from sequential drag interactions. (2) A criterion-guided reprojection pipeline that propagates the inferred criteria to global layout updates. (3) A simulation-based evaluation demonstrating criterion stabilization and layout improvement over existing methods.

2 Related Work

Interactive projection and semantic interaction. DR techniques project high-dimensional data into 2D layouts [6] that users can inspect and manipulate. SI [4, 21] treats direct manipulations of the projection as implicit expressions of analytic intent, which a backing model translates into layout updates. Andromeda [16] adjusts per-dimension weights in Weighted Multi-Dimensional Scaling (WMDS) from SI. DeepSI [2] fine-tunes text embeddings within the interaction loop, and ImageSI [10] extends this paradigm to images by learning embedding weights through inverse MDS on image features. Related work on latent-space editing for images [20] similarly updates representations from user demonstrations. These methods progressively adapt how projections respond to SI, yet the clustering criterion remains implicit in the updated model: the system never surfaces a named criterion, nor can it generalize the inferred intent to images the user has not manipulated.

Language-conditioned image clustering. A complementary line of work treats the clustering criterion as an explicit user-provided natural-language input. IC|TC [9] clusters images under user-provided textual criteria using vision-language models. MultiMaP [24] supports personalized multiple clustering in which users supply brief keywords and LLMs formulate criterion-specific textual contexts. Interactive Semantic Mapping [12] steers projections through natural-language prompts combined with zero-shot classification. More broadly, LLM-augmented visual analytics workflows [26, 18] demonstrate the value of symbolic semantic reasoning in user-facing tools. These approaches give users direct control over *what* to cluster by, but require predefined criteria, limiting their applicability when intent is still being formed.

CriterionSI bridges these gaps, taking only drag actions yet surfacing an explicit, named criterion that emerges through interaction rather than being specified upfront.

3 Method

CriterionSI targets exploratory analytical settings in which the user starts with a tentative sense of how to cluster the image collection, and refines this clustering criterion progressively through drag interactions. Given a collection of images encoded by a frozen vision encoder and projected to 2D via MDS, CriterionSI obtains an initial layout $Y^{(0)}$ on which all subsequent interactions and updates operate. The method then proceeds in three stages (Fig. 2), realized by two components that form a closed loop.

3.1 Interaction capture

At step t, the user drags an anchor image a_t from one canvas position to another. We encode this action as

$$e_t = (a_t,\, N_t^{\text{old}},\, N_t^{\text{new}},\, y_t,\, \tilde{y}_t), \tag{1}$$

where $y_t, \tilde{y}_t$ are the anchor's pre- and post-drag positions and $N_t^{\text{old}}, N_t^{\text{new}}$ are the k-nearest neighborhoods around y_t and $\tilde{y}_t$ on the 2D canvas. From the geometry of the trajectory relative to the two neighborhoods, we classify the interaction into a mode $g_t \in \{\textit{toward}, \textit{away}, \textit{both}, \textit{unclear}\}$, where *both* denotes simultaneous toward and away signals. The mode g_t then feeds the criterion inference and confidence gating stages that follow.

3.2 Criterion-guided Semantic Interaction

Criterion tracker. The criterion tracker (Fig. 2a) makes at most three LLM calls per step: criterion inference, confidence gating, and conditional value mapping. A deterministic state machine then consolidates these per-step hypotheses into a persistent decision.

Criterion inference. Given the anchor image, both neighborhoods, and the interaction mode g_t, the LLM ranks candidate criteria and returns up to three, denoted c_1, c_2, c_3 with associated scores $\gamma_1, \gamma_2, \gamma_3 \in [0,1]$ and $\gamma_1 \geq \gamma_2 \geq \gamma_3$. Because N_t^{new} reflects what the user actively wants to group while N_t^{old} may merely reflect the previous layout, *toward* evidence is weighted more heavily when the two disagree. A small set of reference dimensions, drawn from dataset metadata or summarized by an LLM from a sample of the collection, serves as a shared vocabulary to reduce criterion fragmentation while preserving open-endedness.

Confidence gating. When the top candidates are not clearly separated, a verification pass re-examines whether neighborhood evidence is reliable given the interaction mode g_t, and outputs one of three decisions: *keep*, *revise*, or *defer*. The gating stage also emits a relation label $r_t \in \{\text{support}, \text{conflict}, \text{unclear}\}$ describing how the (possibly revised) top candidate relates to the active criterion, which drives the state transitions below.

Multi-step accumulation. The state machine maintains the active criterion c^*, a support counter n, and a running aggregated score $\hat{\gamma}_c$ for each candidate that is reinforced when an interaction supports c and decayed under conflicting or unclear evidence; transitions consult $\hat{\gamma}_c$ rather than the per-step γ_t to smooth the uncalibrated per-step scores. The transition rules are:

- *Activation*: a candidate becomes the active criterion c^* once it consistently ranks first across the window with $\hat{\gamma}_c \geq \hat{\gamma}_{\min}$.
- *Reinforcement*: r_t=support increments the support counter n and decays the staged challenger's score $\hat{\gamma}_{c_1}$.
- *Challenge and switch*: r_t=conflict buffers the top candidate $c_1 \neq c^*$ as a challenger; the challenger replaces c^* only when $\hat{\gamma}_{c_1} - \hat{\gamma}_{c^*} > \delta$ is sustained across the window.
- *Propagation gate*: when c^* is active and n exceeds a minimum, the LLM is queried to assign each image a value v_i under c^* (Fig. 2b, top); otherwise reprojection updates from drag blending alone, with D^{sem}=$D^{(t-1)}$ unchanged.

Once the tracker produces stable value assignments, the system reshapes the layout through criterion-guided reprojection, drag-constrained blending, and incremental MDS (Fig. 2b). All distances below are computed in the 2D canvas coordinate system of the current layout, ensuring that the subsequent blending operates on a shared scale.

Criterion-guided reprojection. Let $y_i^{(t-1)}$ denote the 2D canvas position of image i after the previous reprojection step, and let $D_{ij}^{(t-1)} = \|y_i^{(t-1)} - y_j^{(t-1)}\|_2$ be the pairwise distance between images i and j in the current layout. For the anchor $i = a_t$, y_t denotes its pre-drag position and $\tilde{y}_t$ denotes its post-drag target. We scale these distances by value agreement under the active criterion:

$$D_{ij}^{\text{sem}} = D_{ij}^{(t-1)} \cdot \begin{cases} 1-\beta, & v_i = v_j, \\ 1+\beta, & v_i \neq v_j, \end{cases} \tag{2}$$

where $\beta \in [0,1)$ controls the shaping strength. Same-value pairs are pulled closer and different-value pairs are pushed apart. When no stable criterion is available, $D^{\text{sem}} = D^{(t-1)}$.

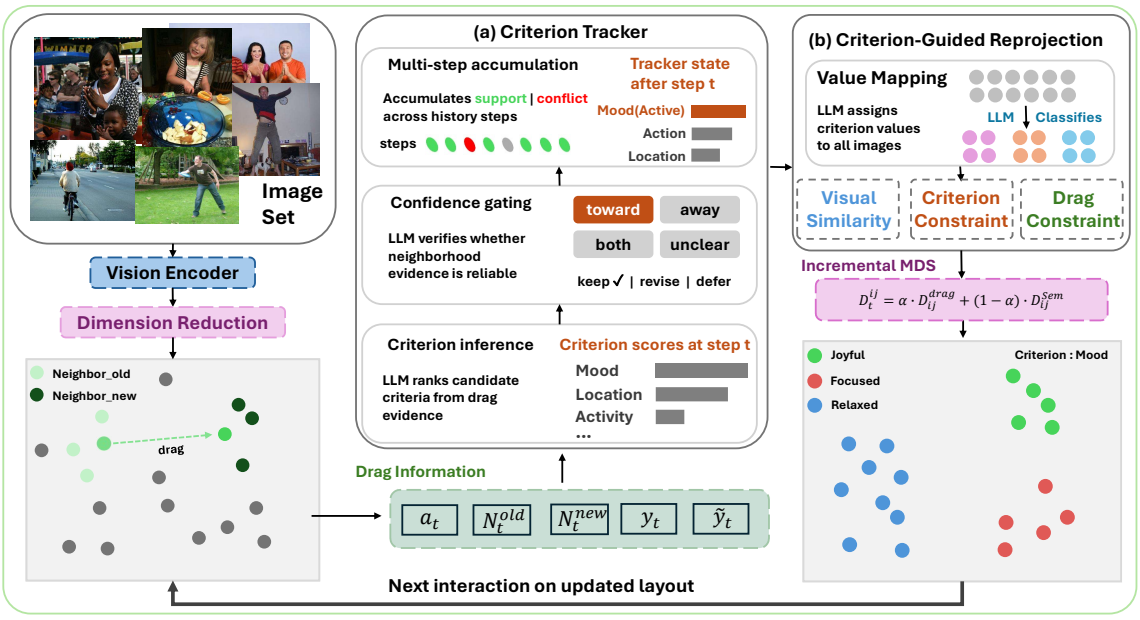

Figure 2: Overview of CriterionSI: (a) Criterion Tracker interprets each drag as partial evidence of a latent clustering criterion. (b) Criterion-Guided Reprojection reshapes pairwise distances and updates the layout through incremental MDS, closing the loop for the next interaction.

Drag-constrained blending. For pairs involving the drag anchor, we compute $D_{ij}^{\text{drag}} = \|\tilde{y}_{a_t} - y_j^{(t-1)}\|_2$, the distance implied by placing the anchor at its target position while neighbors retain their current coordinates. The target distances for these pairs are:

$$D_{ij}^{\text{target}} = \alpha \cdot D_{ij}^{\text{drag}} + (1-\alpha) \cdot D_{ij}^{\text{sem}}, \quad (i,j) \in P_t, \tag{3}$$

where $\alpha \in [0,1]$ controls drag weight and $P_t = \{\{a_t, j\} \mid j \in N_t^{\text{old}} \cup N_t^{\text{new}}\}$ is the set of anchor-neighbor pairs. For all other pairs, $D_{ij}^{\text{target}} = D_{ij}^{\text{sem}}$.

Incremental MDS. The updated layout $Y^{(t)}$ is obtained by minimizing stress $\sum_{i<j}(\|y_i - y_j\| - D_{ij}^{\text{target}})^2$ from the current positions, with the $K \geq 0$ most recent drag anchors pinned at their post-drag targets. This sliding-window pinning ensures that early drags do not dominate the optimization as the interaction progresses.

4 Usage Scenario

We illustrate how the tracker evolves through a simulated session on the *Mood* task (clustering the Action40 collection by mood). All silhouette values in this section are from that one run, and experiment settings follow Sec. 5.1.

Initial projection. The initial layout is arranged by visual similarity, leaving the three mode classes spatially intermixed (Fig. 3a).

Ambiguous early drags (steps 1–5). In the first few drags, the LLM returns near-tied candidate criteria across steps. After aggregation over the sliding window, no candidate reaches the activation threshold, so the tracker remains uncommitted. The layout shows no clear cluster structure ($s_5 = -0.03$):

> *"selected_criterion:* ***social context****; selected_value:* ***solitary*** *(0.72); alternatives:* ***mood/joyful*** *(0.60),* ***age/children*** *(0.60)."* [step 5]

Criterion stabilization (step 6). The next *toward* drag groups a man-jumping image with a bubble-blowing scene; the LLM disambiguates, and *mood*'s aggregated score $\hat{\gamma}_{mood} = 0.64$ crosses the activation threshold over the sliding window, while the per-step score reaches 0.90 at this step.

> *"criterion_action:****activate****; selected_criterion:* ***Mood*** *(0.90). Anchor and new group consistently depict active joy, playfulness, or enthusiastic enjoyment."* [step 6]

Silhouette turns positive at activation ($s_6 = 0.04$) and grows incrementally as the criterion-guided reprojection reshapes the layout over subsequent steps, reaching $s_{20} = 0.35$ (Fig. 3f).

Refinement under momentary conflict (steps 7–40). Supporting *toward* drags re-confirm *mood*; even a step-20 challenge nominating *Age Extremes* (0.95) fails the state machine's switch criterion, since its lead over mood does not persist across the sliding window. At step 28 a bubble-blowing anchor briefly prompts the LLM to rank *Recreational Activity* above *mood*:

> *"active_criterion_check:* ***conflict****; selected_criterion:* ***Recreational Activity*** *(0.95)."* [step 28]

Because the state machine evaluates accumulated support across multiple steps rather than a single peak, this one-step spike fails to meet the switch threshold; the cumulative evidence for *mood* from prior steps still dominates. Value mapping therefore continues under $c^* = mood$, and the silhouette holds at $s_{40} = 0.59$ with the three mood clusters cleanly resolved (Fig. 3g).

5 Evaluation

We evaluate CriterionSI through a simulation-based quantitative comparison and a representative usage scenario on the *Mood* task.

Implementation details. Images are encoded with EVA-CLIP ViT-G/14 [17] and all LLM calls use Gemini 2.5 Flash. We use neighborhood k=3, window W=3, activation $\hat{\gamma}_{\min}$=0.5, switch margin δ=0.15, shaping β=0.3, drag-blend α=0.7, and anchor-pinning K=5. Prompts, datasets, more implementation details, and more usage scenarios are in the supplemental material.

5.1 Experimental Setup

Simulation engine. We simulate sequential user behavior with a goal-directed drag generator. At each step, an anchor image is selected from the unvisited pool and moved toward the centroid of the same-label images that have already been positioned. The simulator

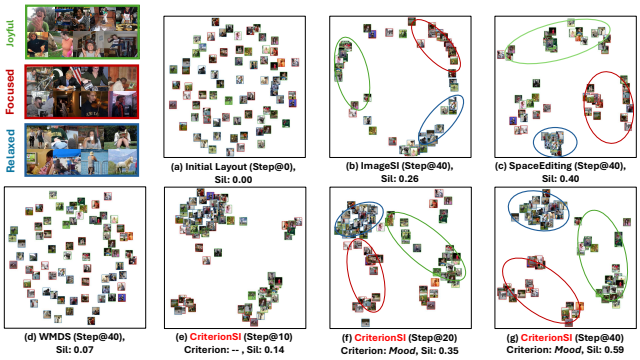

Figure 3: DR plots after interaction on the *Mood* task. (same as Sec. 4) (a) Initial layout; (b-d) ImageSI, SpaceEditing, WMDS at step 40; (e-g) CriterionSI (Gemini cls) at steps 10, 20, and 40.

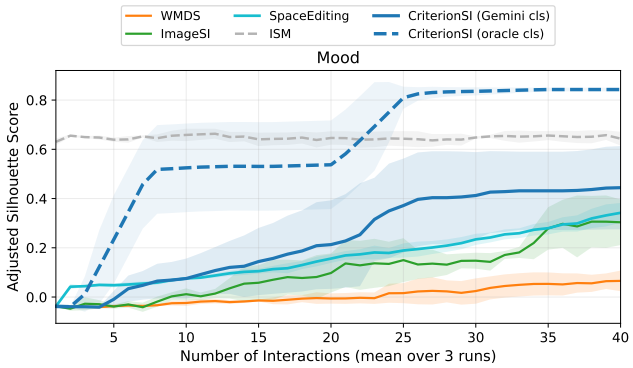

Figure 4: Per-step adjusted Silhouette scores on the *Mood* task; shaded bands show variance across simulated sessions.

uses ground-truth labels only for drag target generation; the tracker receives only the drag event and neighborhood images, with no label access. No purity constraints are enforced on neighborhoods, so the method is exposed to noisy and ambiguous evidence.

Dataset and comparison methods. We use a 72-image subset of Action40 [23] and cluster it into the three Mood classes *Joyful*, *Focused*, and *Relaxed* defined by IC|TC [9]. We focus on Mood because it is the most challenging of the Action40 dimensions: *mood* is weakly aligned with visual-appearance distances. All methods share the same EVA-CLIP [17] initial representation and projection; they differ only in the signals used to update the layout, summarized in Tab. 1. ISM [12] refers to the same CriterionSI pipeline without drag interaction: it receives the ground-truth criterion upfront and shares oracle value assignments with CriterionSI (oracle cls). The resulting comparison isolates drag-driven criterion tracking and reprojection. We additionally report two variants of our method: CriterionSI (Gemini cls) is the fully automatic pipeline in which Gemini 2.5 Flash performs both criterion inference and value classification; CriterionSI (oracle cls) uses ground-truth value assignments to isolate criterion-tracking quality from classification noise.

5.2 Quantitative Results

Fig. 4 shows the per-step silhouette trajectory; all numbers below are averaged over three runs. Starting from $s_0 \approx 0.00$, CriterionSI (Gemini cls) reaches $s_{40} = 0.44$, substantially outperforming other baselines under the same interaction budget. Across our tasks, the criterion commits within roughly 10% of the dataset size, after which value mapping propagates it to all images at once. WMDS struggles to rise above $s \approx 0.1$, since its reweighting can only reshape visual similarity and cannot separate images that look alike but belong to different moods. ImageSI and SpaceEditing steadily improve but level off in the mid-range, because their weight adaptations struggle to learn discriminative features from drag interactions when the clustering criterion is not visually salient.

Notably, CriterionSI (oracle cls) reaches $s_{40} = 0.84$, exceeding ISM ($s_{40} = 0.65$) by 0.19. Since the value mapping is controlled, this gap shows that drag-driven tracking and reprojection contribute a local grouping signal that pure criterion-based clustering cannot provide. The Gemini cls shortfall stems from classification noise on *mood* labels, leaving classification fidelity as the bottleneck.

Table 1: Compared methods by the signals used to update the layout.

Method	Drag	Visual	Language	Criterion
WMDS [16]	✓			
ImageSI [10]	✓	✓		
SpaceEditing [20]	✓	✓		
CriterionSI (ours)	✓	✓	✓	inferred
ISM [12]		✓	✓	predefined

6 Discussion

Rethinking "correct" criterion recovery in SI. CriterionSI does not always converge on the exact target label. When the target is *Action*, it may stabilize on *body posture*, since dynamic movements and static postures yield a largely compatible partition. This is not an isolated case: semantically related criteria such as posture and action, or mood and expression, often carve the data along near-parallel boundaries, making criterion tracking a many-to-one problem at the partition level. We observed no clear performance degradation under near-miss convergence; surfacing a neighboring criterion can be informative, revealing semantic dimensions the user had not explicitly considered. We therefore view CriterionSI less as a system that must guess the right label, and more as one that commits to a coherent partition the user can accept, reject, or refine.

Co-constructed intent through divergent reasoning. Across our sessions, LLMs often propose criteria the user had not articulated, and occasionally reason along dimensions that do not match the user's initial target. Rather than treating these divergences as inference failures, CriterionSI treats them as hypotheses to be tested against continued interaction. Partial human signals accumulate through drags, LLM hypotheses that fail to find support decay, and the criterion that stabilizes is one both parties have implicitly agreed on through different modes of reasoning. Intent is thus neither extracted from the user nor generated by the model, but co-constructed: the state machine is where human evidence and LLM hypotheses negotiate, and the resulting criterion is often one neither party would have written down at step one.

Limitations. Because following a criterion requires a fraction of the images (Sec. 5.2), interaction cost grows with collection size and dataset complexity; proactively proposing the criterion once enough evidence has accumulated could curb this. The evaluation also relies on a simulated, goal-consistent user that excludes human error and mid-session reconsideration; whether these results hold for real users is an open question we leave to future work.

Conclusion and Future Work. In this work, we introduced CriterionSI, a drag-based approach that couples user interaction with LLM-driven criterion inference for image clustering. Simulation experiments showed CriterionSI outperforms prior interactive baselines and rivals settings in which the criterion is specified upfront. For future work, we see CriterionSI as a step toward agentic visualization that reasons alongside users, with proactive criterion suggestion, multi-agent negotiation, and persistent cross-session memory as natural next directions.

REFERENCES

[1] Y. Bian, R. Faust, and C. North. NeuralSI: Neural design of semantic interaction for interactive deep learning. *arXiv preprint arXiv:2402.17178*, 2024. doi: 10.48550/arXiv.2402.17178 1

[2] Y. Bian and C. North. DeepSI: Interactive deep learning for semantic interaction. In *Proceedings of the 26th International Conference on Intelligent User Interfaces*, pp. 197–207. ACM, 2021. doi: 10.1145/3397481.3450670 1, 2

[3] M. Cavallo and Ç. Demiralp. Clustrophile 2: Guided visual clustering analysis. *IEEE Transactions on Visualization and Computer Graphics*, 25(1):267–276, 2019. doi: 10.1109/TVCG.2018.2864477 1

[4] A. Endert, P. Fiaux, and C. North. Semantic interaction for sensemaking: inferring analytical reasoning for model steering. *IEEE Transactions on Visualization and Computer Graphics*, 18(12):2879–2888, 2012. doi: 10.1109/TVCG.2012.260 1, 2

[5] M. Espadoto, R. M. Martins, A. Kerren, N. S. Hirata, and A. C. Telea. Toward a quantitative survey of dimension reduction techniques. *IEEE Transactions on Visualization and Computer Graphics*, 27(3):2153–2173, 2021. doi: 10.1109/TVCG.2019.2944182 1

[6] A. K. Jain. Data clustering: 50 years beyond k-means. *Pattern Recognition Letters*, 31(8):651–666, 2010. doi: 10.1016/j.patrec.2009.09.011 1, 2

[7] D. A. Keim, F. Mansmann, J. Schneidewind, J. Thomas, and H. Ziegler. Visual analytics: Scope and challenges. In *Visual data mining: Theory, techniques and tools for visual analytics*, pp. 76–90. Springer, 2008. doi: 10.1007/978-3-540-71080-6_6 1

[8] B. C. Kwon, B. Eysenbach, J. Verma, K. Ng, C. De Filippi, W. F. Stewart, and A. Perer. Clustervision: Visual supervision of unsupervised clustering. *IEEE Transactions on Visualization and Computer Graphics*, 24(1):142–151, 2018. doi: 10.1109/TVCG.2017.2745085 1

[9] S. Kwon, J. Park, M. Kim, J. Cho, E. K. Ryu, and K. Lee. Image clustering conditioned on text criteria. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2024. 1, 2, 4

[10] J. Lin, R. Faust, and C. North. ImageSI: Semantic interaction for deep learning image projections. In *2024 IEEE Visualization and Visual Analytics (VIS)*, pp. 91–95. IEEE, 2024. doi: 10.1109/VIS55277.2024.00026 1, 2, 4

[11] L. G. Nonato and M. Aupetit. Multidimensional projection for visual analytics: Linking techniques with distortions, tasks, and layout enrichment. *IEEE Transactions on Visualization and Computer Graphics*, 25(8):2650–2673, 2019. doi: 10.1109/TVCG.2018.2846735 1

[12] A. A. Oliveira, M. Espadoto, R. Hirata Jr, R. M. Cesar Jr, and A. C. Telea. Creating user-steerable projections with interactive semantic mapping. *arXiv preprint arXiv:2506.15479*, 2025. doi: 10.48550/arXiv.2506.15479 1, 2, 4

[13] P. Pirolli and S. Card. The Sensemaking Process and Leverage Points for Analyst Technology as Identified through Cognitive Task Analysis. In *Proceedings of the International Conference on Intelligence Analysis*, pp. 2–4. McLean, VA, USA, 2005. 1

[14] D. Sacha, A. Stoffel, F. Stoffel, B. C. Kwon, G. Ellis, and D. A. Keim. Knowledge generation model for visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 20(12):1604–1613, 2014. doi: 10.1109/TVCG.2014.2346481 1

[15] D. Sacha, L. Zhang, M. Sedlmair, J. A. Lee, J. Peltonen, D. Weiskopf, S. C. North, and D. A. Keim. Visual interaction with dimensionality reduction: A structured literature analysis. *IEEE Transactions on Visualization and Computer Graphics*, 23(1):241–250, 2017. doi: 10.1109/TVCG.2016.2598495 1

[16] J. Z. Self, M. Dowling, J. Wenskovitch, I. Crandell, M. Wang, L. House, S. Leman, and C. North. Observation-level and parametric interaction for high-dimensional data analysis. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 8(2):1–36, 2018. doi: 10.1145/3158230 2, 4

[17] Q. Sun, Y. Fang, L. Wu, X. Wang, and Y. Cao. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*, 2023. doi: 10.48550/arXiv.2303.15389 3, 4

[18] X. Tang, I. Tahmid, E. Krokos, K. Whitley, X. Wang, and C. North. Semantic prompting: Agentic incremental narrative refinement through spatial semantic interaction. *arXiv preprint arXiv:2604.19971*, 2026. doi: 10.48550/arXiv.2604.19971 2

[19] C. Ware. *Information visualization: perception for design*. Morgan Kaufmann, 2019. 1

[20] J. Wei, D. Xia, H. Xie, C.-M. Chang, C. Li, and X. Yang. SpaceEditing: A latent space editing interface for integrating human knowledge into deep neural networks. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pp. 489–503. ACM, 2024. doi: 10.1145/3640543.3645211 1, 2, 4

[21] J. Wenskovitch, M. Dowling, and C. North. Toward addressing ambiguous interactions and inferring user intent with dimension reduction and clustering combinations in visual analytics. *ACM Transactions on Interactive Intelligent Systems*, 14(1):1–35, 2024. doi: 10.1145/3588565 1, 2

[22] J. Xia, Y. Zhang, J. Song, Y. Chen, Y. Wang, and S. Liu. Revisiting dimensionality reduction techniques for visual cluster analysis: An empirical study. *IEEE Transactions on Visualization and Computer Graphics*, 28(1):529–539, 2022. doi: 10.1109/TVCG.2021.3114694 1

[23] B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. Guibas, and L. Fei-Fei. Human action recognition by learning bases of action attributes and parts. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1331–1338. IEEE, 2011. doi: 10.1109/ICCV.2011.6126386 4

[24] J. Yao, Q. Qian, and J. Hu. Multi-modal proxy learning towards personalized visual multiple clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14066–14075. IEEE, 2024. doi: 10.1109/CVPR52733.2024.01334 2

[25] J. Zhang, J. Huang, S. Jin, and S. Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(8):5625–5644, 2024. doi: 10.1109/TPAMI.2024.3369699 1

[26] Y. Zhao, Y. Zhang, Y. Zhang, X. Zhao, J. Wang, Z. Shao, C. Turkay, and S. Chen. LEVA: Using large language models to enhance visual analytics. *IEEE Transactions on Visualization and Computer Graphics*, 31(3):1830–1847, 2025. doi: 10.1109/TVCG.2024.3368060 1, 2