Supplemental for DKMap

July 2, 2025

1 Additional Quantitative Results

In this section we show some additional quantitative results, including in-sample mapping accuracy (Table 1), neighborhood preservation metric (Table 2), and running time comparison on large-scale embeddings (Table 1). We also show unimodal embedding metric mapping results in Figure 2.

Table 1: in-sample mapping accuracy.

Method	CLIPScore			HPSv2			MPS			PickScore		
	mae	*mape*	*rmse*	*mae*	*mape*	*rmse*	*mae*	*mape*	*rmse*	*mae*	*mape*	*rmse*
PCA (KDE)	3.049	10.450	3.912	1.403	5.093	1.796	2.915	352.999	3.693	1.144	5.572	1.447
PCA (IDW)	2.925	10.033	3.915	1.324	4.806	1.772	2.655	388.061	3.558	1.095	5.336	1.446
t-SNE (KDE)	2.986	10.187	3.831	1.343	4.875	1.730	2.859	475.663	3.605	1.086	5.287	1.378
t-SNE (IDW)	2.215	7.547	3.083	1.069	3.879	1.460	1.462	138.969	2.044	0.807	3.916	1.097
UMAP (KDE)	3.032	10.354	3.886	1.471	5.341	1.880	3.434	634.234	4.337	1.185	5.788	1.508
UMAP (IDW)	2.432	8.267	3.329	1.100	3.991	1.500	1.591	134.456	2.291	0.845	4.101	1.145
Neuro-Visualizer	9.404	29.228	10.761	3.444	12.068	4.034	3.747	155.654	4.646	2.147	10.038	2.576
ModalChorus	3.418	11.791	4.348	1.529	5.541	1.948	3.695	641.665	4.644	1.177	5.716	1.491
DKMap (w/o KR)	2.955	10.103	3.791	1.307	4.747	1.684	2.701	412.932	3.416	1.083	5.269	1.373
DKMap	1.455	4.821	1.892	0.758	2.735	0.993	0.721	106.826	0.931	0.704	3.435	0.920

Table 2: Neighborhood Trustworthiness metrics for visualization of HPSv2 embeddings on Pick-a-pic.

	PCA	**t-SNE**	**UMAP**	**DKMap**
n=20	0.7167	0.8692	0.6204	0.8325
n=30	0.7141	0.8440	0.5980	0.8286

Table 3: Evaluation of scatterplot quality by $nRMSE \downarrow$ for CLIPScore.

	PCA	**t-SNE**	**UMAP**	**DKMap**
k=5	1.011	0.947	0.937	0.889
k=10	0.970	0.907	0.899	0.850
k=20	0.948	0.889	0.882	0.831

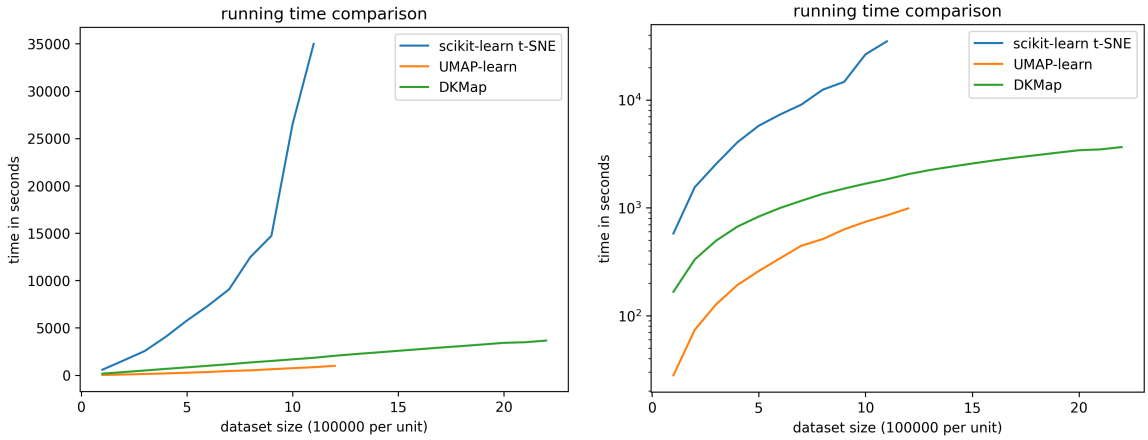
2 User Study

To evaluate the effectiveness of our DR visualization technique in supporting alignment understanding and distribution interpretation, we conduct a user study on 20 users aged between 21 and 40 (9 females, 11 males), where we ask them to compare the different visualization methods on CLIPScore embedding alignment on four dimensions: 1) map accuracy (how well the contour map accurately reflect the distribution in raw scatterplot), 2) scatterplot clarity (how well the scatterplot mode alone can clearly reveal the distribution), 3) map smoothness and 4) map detail. We find that our method consistently outperform other methods in all aspects, which aligns with our quantitative experiments.

To further assess whether our system supports users in generating insights and making decisions, we designed a follow-up questionnaire, involving 6 participants (2 females, 4 males) from the original 20 users who participated in our earlier evaluation. These participants were invited for an in-depth evaluation combining both structured Likert-scale items and open-ended questions. Most participants found the interaction between the Keyword Distribution View and the Overview to be helpful and engaging,

Table 4: Evaluation of CLIPScore mapping accuracy on ImageRewardDB dataset.

	PCA (KDE)	PCA (IDW)	t-SNE (KDE)	t-SNE (IDW)	UMAP (KDE)	UMAP (IDW)	DKMap
MAE	2.659	2.745	2.696	2.451	2.722	2.593	1.981
$MAPE$	8.645	8.919	8.719	7.796	8.786	8.254	6.386
$RMSE$	3.407	3.533	3.422	3.116	3.456	3.326	2.564



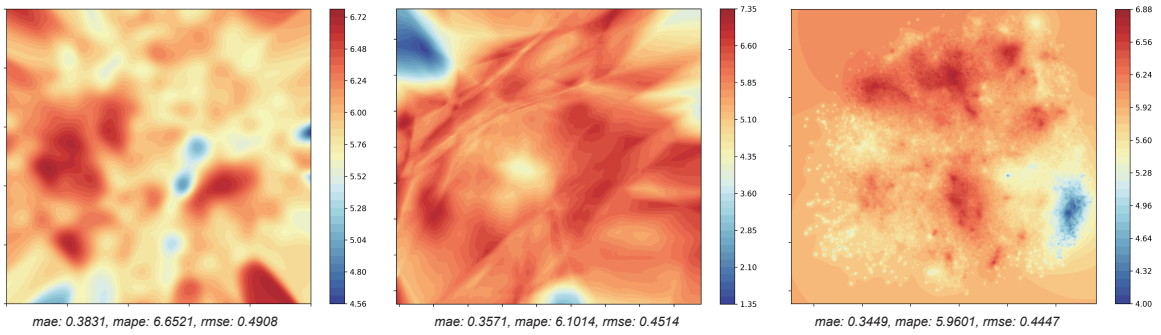
(a) Linear scale running time (b) Log scale running time

Figure 1: Time comparison with two commonly used packages: scikit-learn t-SNE and umap-learn. umap-learn can achieve linear speed but takes up large memory, bursting 80GB memory after 1.2 million, while scikit-learn t-SNE has fast growing time complexity. Our method can also achieve linear complexity and go well beyond 2 million (not the limit of our method; it can be even larger but we do not have time running more experiments).

as it is enabled by filtering keywords according to different similarity levels (P1, P3, P4, P5). However, several participants also pointed out that the quality of the recommended keywords could be further improved. For example, P5 suggested generating *"more expressive or meaningful keywords"* to better capture subtle differences in visual style or subject matter. P3 mentioned that *While some keywords were relevant, others appeared too generic or lacked clear visual meaning.* Beyond keyword feedback, participants also shared their thoughts on the Overview. All participants noted that the Overview view effectively visualized the vision-language alignment patterns. They particularly appreciated its usefulness in comparing models, where shifts or overlaps in clusters became visually apparent. P4 emphasized that *"It's easy to tell how two models perform differently just by looking at the overview. It's very efficient."* Two participants (P3, P5) further suggested incorporating exploration guidance to support misalignment analysis. For instance, they proposed showing a brief summary of the dataset's overall alignment distribution when entering the Overview, or enabling users to manually select a keyword and retrieve all relevant images directly. In one typical exploration scenario, a participant observed that images generated in the style of certain artists consistently exhibited low vision-language alignment. This was often attributed to the abstract or stylized nature of these artworks, which made it harder for the model to establish a clear semantic match with the accompanying text.

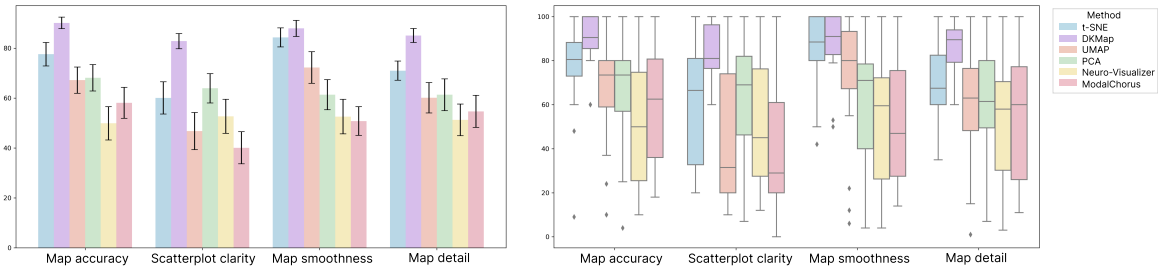
3 Additional Results

References



(a) traditional DR (t-SNE) (b) autoencoer (Neuro-Visualizer) (c) DKMap

Figure 2: DKMap can also more accurately visualize unimodal Aesthetic Score compared to traditional DR. In addition, the problems of auto-encoder is less serious with unimodal embeddings like Aesthetic Score.



(a) User study results (boxplot) (b) User study results (bar chart)

Figure 3: User study rating results in bar chart and box plot.

No.	Question	Score Distribution	Average
Q1	The Overview helps me quickly identify potential misalignment.	2 4	4.67
Q2	The Overlay mode effectively helps me identify important distribution patterns while allowing detailed examination of individual data points.	2 4	4.67
Q3	The Keyword Distribution View helps me interpret the semantic structure and topic variations within the selected data group.	1 5	4.83
Q4	The Instance View allows me to efficiently examine individual samples and explore semantically similar ones.	3 3	4.50
Q5	It is easy to learn how to use the system.	1 5	4.67
Q6	The interaction between the Keyword Distribution View and the Overview is smooth.	2 4	4.67
Q7	The interface is intuitive and easy to navigate.	3 3	4.50
Q8	I would use this system again.	2 4	4.67

1 (Strongly Disagree) 2 3 4 5 (Strongly Agree)

Figure 4: The results of the questionnaire regarding the effectiveness and usability of the Web-based system.

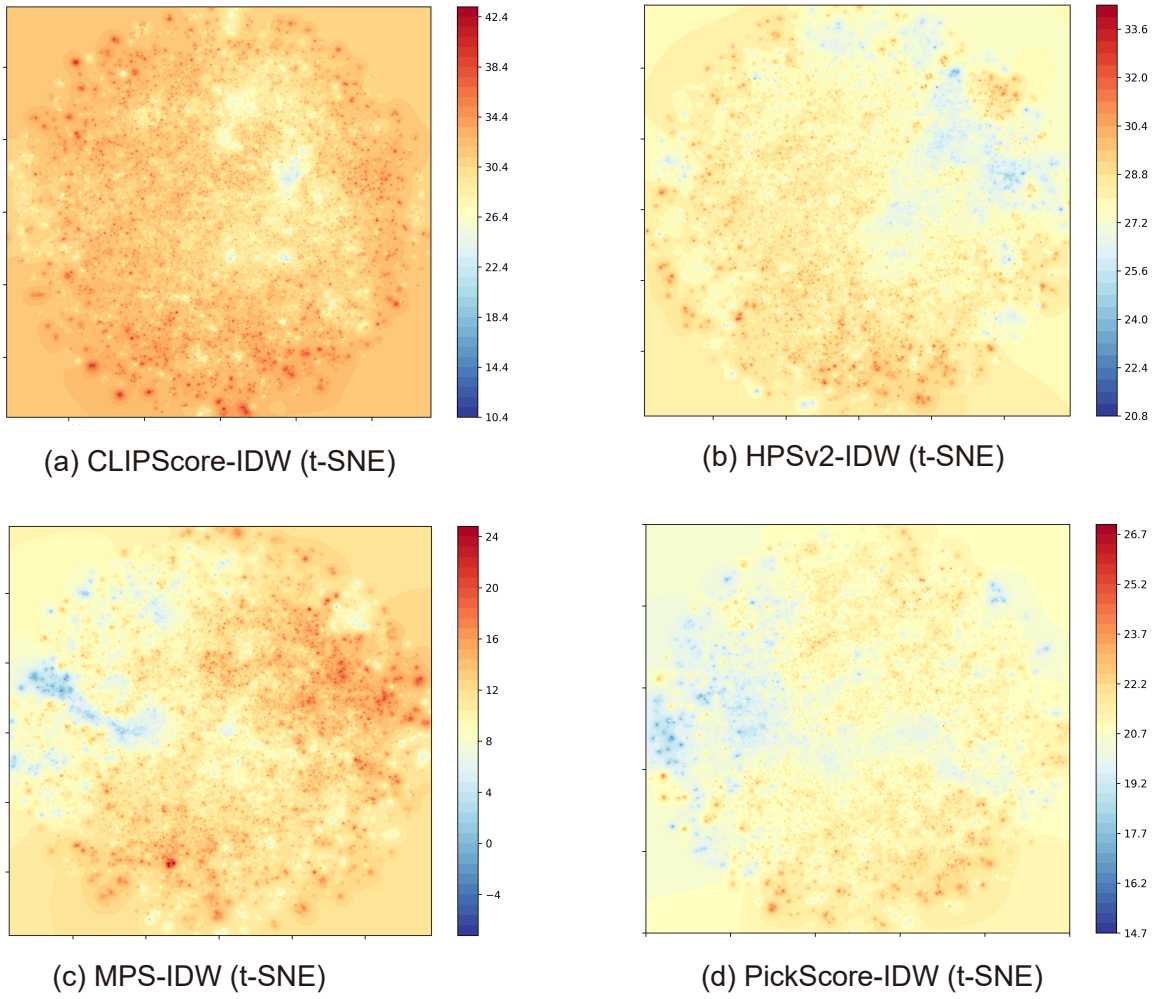


Figure 5: Examples of IDW contour map.

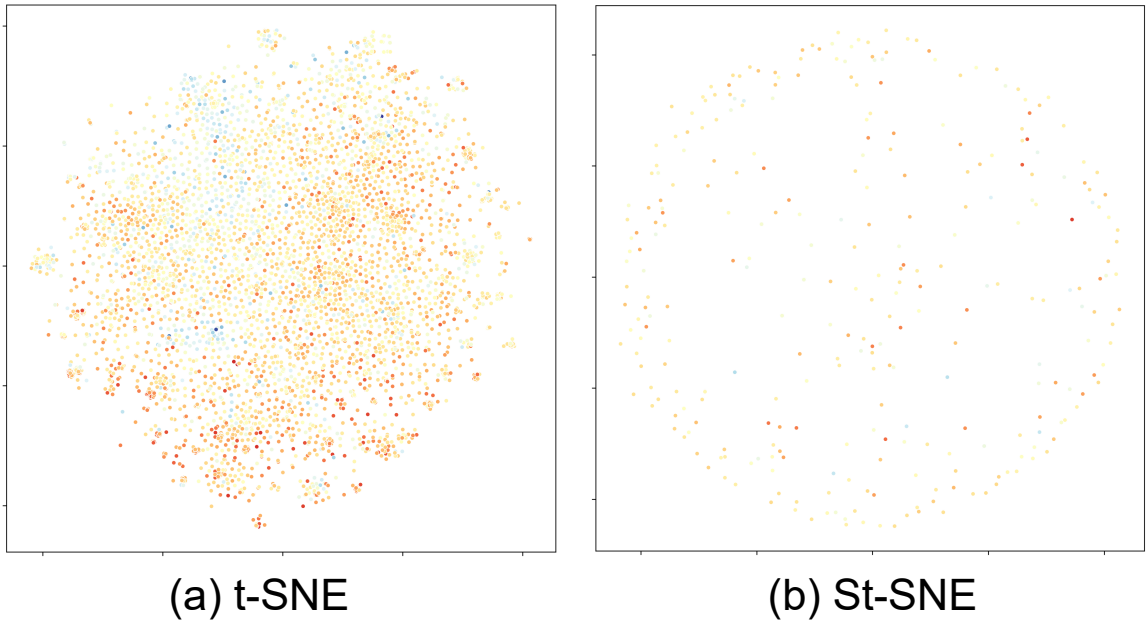


Figure 6: St-SNE results on 10000 sampled points from HPSv2 embeddings of Pick-a-pic dataset. St-SNE causes collapse in the projection with most points overlapping in (b) compared to (a) original t-SNE that can still reveal the 10000 points.