```
{
    "ID": "2401.13641",
    "title": "How Good is ChatGPT at Face Biometrics? ...",
    "authors": ["...", "..."],
    "published": "...",
    "subjects": {
        "arXiv": ["...", "..."], "ACM": ["...", "..."], "MSC": ["...", "..."]
    },
    "comment": "...",
    "journal_ref": "IEEE Access, February 2024",
    "conference": "...",
    "DOI": ["https://doi.org/10.48550/arXiv.2401.13641", "..."],
    "abstract": "Large Language Models (LLMs) such as GPT developed ..."
    "graphical_abstract": {
        "ID": "2401.13641_GA",
        "type": "Reused",
        "path": ["..."],
        "components": ["2401.13641_F1"],
        "caption": "...",
    },
    "teaser": ["2401.13641_F1"],
    "sections": {
        "ID": "2401.13641_S1",
        "title": "<TAG> 1 </TAG> Introduction",
        "body": "...",
        "subsections": {...},
        "figures": ["2401.13641_F1"],
    }, {...}
    "figures": {
        "ID": "2401.13641_F1",
        "caption": "<TAG> Fig. 1 </TAG> Graphical representation of ..."
        "path": ["..."],
        "subfigures": {...}
    }, {...}
}
```


Listing 1. Example data in SciGA-145k.[6]

A. Dataset Structure

Below, we describe the structure of SciGA-145k, focusing on how metadata, textual content, and visual content are represented.

Metadata and Textual Data. SciGA-145k includes metadata and textual content extracted from scientific papers. Each entry is formatted in JSON and contains standard metadata fields (e.g., title, authors, abstract, research fields, DOI) as well as structured content such as section hierarchy and figure composition, including subfigures and captions. This rich structural representation supports fine-grained figure-level retrieval and content analysis. A partial example of the JSON structure is shown in Listing 1.

Visual Data. SciGA-145k includes figure data extracted from scientific papers. These visual assets are provided in PNG format for static figures and MP4 format for video-based GAs. While most in-paper figures were successfully extracted from the HTML-rendered versions of arXiv papers, approximately 50k instances ($\sim$ 4–5%) exhibited rendering failures, either appearing as blank placeholders or visibly corrupted images. These failures were typically caused by complex TEX figure commands (e.g., `\includegraphics` with `page`, `trim`, or `clip`) that reference partial PDF content. To ensure dataset completeness, such figures were manually extracted from the original TEX source files. In contrast,

[6] arXiv ID: 2401.13641

GAs were manually collected from open-access journal versions identified via arXiv metadata (e.g., related DOI, journal reference, or author comments). However, only a small subset of papers include GAs, as most journals do not require them or restrict their reuse due to license limitations. Moreover, available GAs are often limited to specific journals and disciplines, making them unsuitable for constructing large-scale, balanced benchmarks, which is a limitation also observed in prior studies that typically rely on fewer than 100 samples. To enable broader domain coverage and consistent empirical evaluation, we adopt in-paper teaser figures as scalable proxies for GAs, automatically labeled based on their position and layout. Specifically, we heuristically identify teasers as figures that are cited as *Figure 1* in the *Introduction* and appear: 1) on the first page in full-column width or in the upper-right position, or 2) on the second page in full-column width.

B. Content Statistics

This section presents summary statistics of SciGA-145k, covering research field distribution, content structure, and GA creation trends. These analyses complement the main paper and highlight the dataset's diversity.

Research Field Coverage. Research fields in SciGA-145k are primarily classified using arXiv's hierarchical category system. In addition, some papers include author-supplied labels from the ACM Computing Classification System (ACM-CCS) [7] and the Mathematics Subject Classification 2022 (MSC2022),[8] extracted from arXiv metadata or full-text content. These entries were often unstructured or inconsistent, so we curated and normalized them using a combination of rule-based preprocessing and manual refinement (e.g., resolving casing and separator inconsistencies, removing label prefixes, assigning primary/secondary roles based on position or formatting, and identifying the taxonomy).

Fig. 9 shows the distribution of primary arXiv categories, with computer science (`cs`, 39.6%) and mathematics (`math`, 16.6%) as the largest categories, followed by astrophysics (`astro-ph`, 8.1%) and condensed matter physics (`cond-mat`, 6.4%). This distribution demonstrates the dataset's broad domain coverage and suitability for cross-disciplinary analysis.

Textual and Visual Statistics. Fig. 10 summarizes token counts for titles, abstracts, full texts, captions, and the number of figures per paper across research fields. Titles show little variation across domains, while abstracts tend to be longer in `astro-ph`, reflecting the field's more descriptive writing style. Full texts are notably lengthy in economics (`econ`), whereas `cs` and electrical engineering and systems science (`eess`) papers are generally shorter, following concise conference-oriented formats compared to journal-style publications. Each paper includes on average 6.16 ± 5.86 figures (7.92 ± 10.45 including subfigures), with astrophysics papers often exceeding 10 figures, compared to about 4 for mathematics, highlighting disciplinary differences in visual representation. Captions also vary widely, with experimental sciences frequently featuring more detailed and informative figure descriptions

Table 4. Distributional distances between GAs, teasers, and regular in-paper figures in CLIP space. GA–Teaser distances are extremely small under both CMMD and FD, indicating that the two figure types share nearly identical embedding distributions.

	CMMD↓	FD↓
GA–Teaser	0.078	0.486
GA–*Regular In-paper Figure*	0.284	3.888
Teaser–*Regular In-paper Figure*	0.128	1.974

Examining GA Origins and Their Relation to Teasers. Although GAs and teasers originate from different sources (journal-submitted vs. in-paper), they serve the same communicative role and are functionally indistinguishable. To substantiate this claim, we examine whether any cultural or distributional distinctions between GAs and teasers exist. First, we analyze how authors create their GAs by categorizing each GA into the following three types: 1) *Original*: newly created GAs without reusing any in-paper figures; 2) *Reuse*: GAs directly copied from figures in the paper without modifications; and 3) *Modified*: GAs created by combining or modifying in-paper figures. Annotations were manually performed by one author using deterministic criteria based on layout similarity, content overlap, and evidence of cropping or compositional edits, and all labels were reviewed with two co-authors. Among the 309 GAs, 20.9% were categorized as *Original*, 64.5% as *Reused*, and 14.5% as *Modified*, consistent with prior reports on GA creation patterns [59]. These statistics indicate that authors themselves do not meaningfully distinguish between GAs and in-paper teasers. Second, we analyze whether this cultural pattern is reflected in their embedding distributions. Tab. 4 summarizes the CMMD [17] and Fréchet distances (FD) [51] between GAs, teasers, and regular in-paper figures in CLIP space. These metrics quantify the distributional discrepancy among figure types, and both clearly show that the GA–Teaser distance is extremely small, indicating that their embedding distributions are nearly identical. When distances are measured within each research field, the GA–teaser distance becomes even smaller (average CMMD: 0.072), whereas GA–teaser distances across different fields are substantially larger (0.428). These results suggest that the communicative role of a figure varies along two factors: 1) whether it is a

[7] https://dl.acm.org/ccs
[8] https://doi.org/10.4171/news/115/2

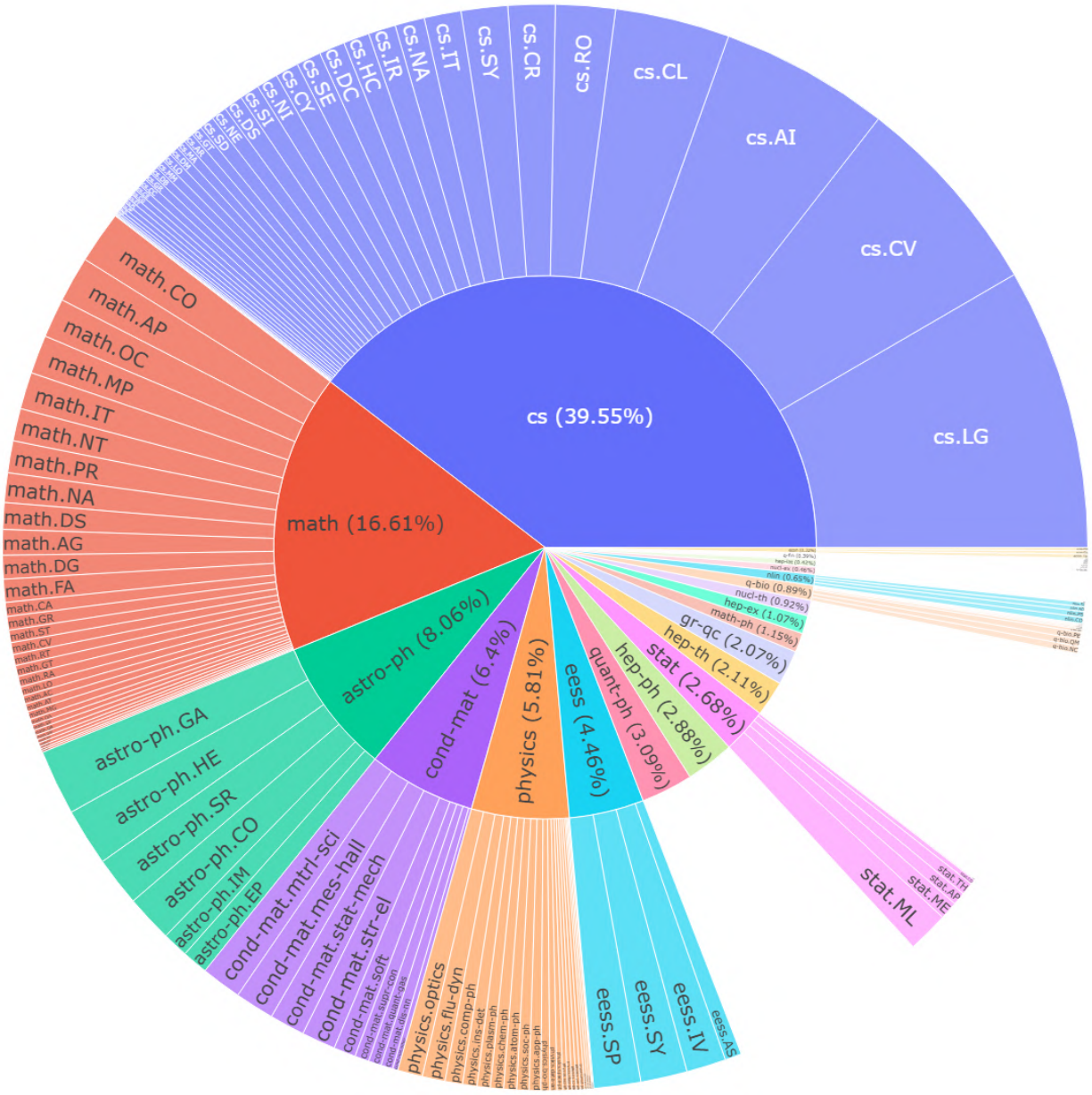
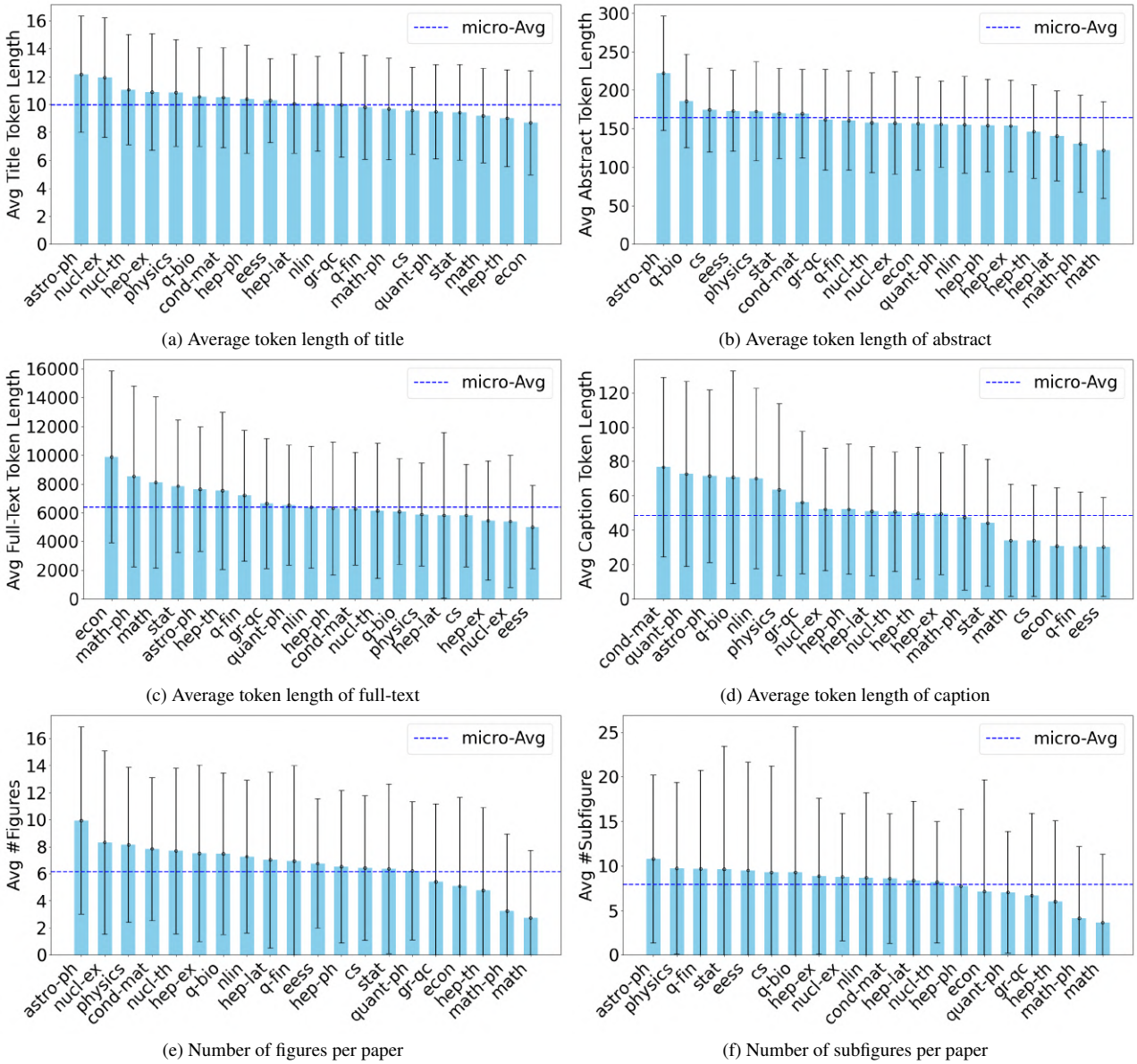


Figure 9. Distribution of research fields of papers included in SciGA-145k.

summary-style figure (irrespective of being GA or teaser) or an ordinary figure; and 2) which research field it belongs to. Overall, our observations indicate that teasers and GAs can be regarded as fundamentally equivalent summary-style figures.

C. Threshold Sensitivity of CAR

We analyze the sensitivity of the confidence adjustment term $\mathcal{C}(P, k)$ in CAR@k with respect to the threshold parameter h. For this analysis, we reformulate it as a fraction of the maximum entropy, with $h = \alpha H_{\max}(P)$, where $\alpha \in [0, 1]$ adjusts the sensitivity of confidence evaluation. To evaluate the effect of α, we measured the distribution of CAR@5 scores across test queries under different threshold values (see Fig. 11 for score distributions and Fig. 12 for their means and standard deviations). When α is too low, $\mathcal{C}(P, k)$ tends to underestimate confidence even when the relevance distribution is sharply peaked, leading to compressed CAR@k scores near the lower end of the $[0, 1]$ range. Conversely, excessively high α values yield overconfident $\mathcal{C}(P, k)$ es-



(a) Average token length of title

(b) Average token length of abstract

(c) Average token length of full-text

(d) Average token length of caption

(e) Number of figures per paper

(f) Number of subfigures per paper

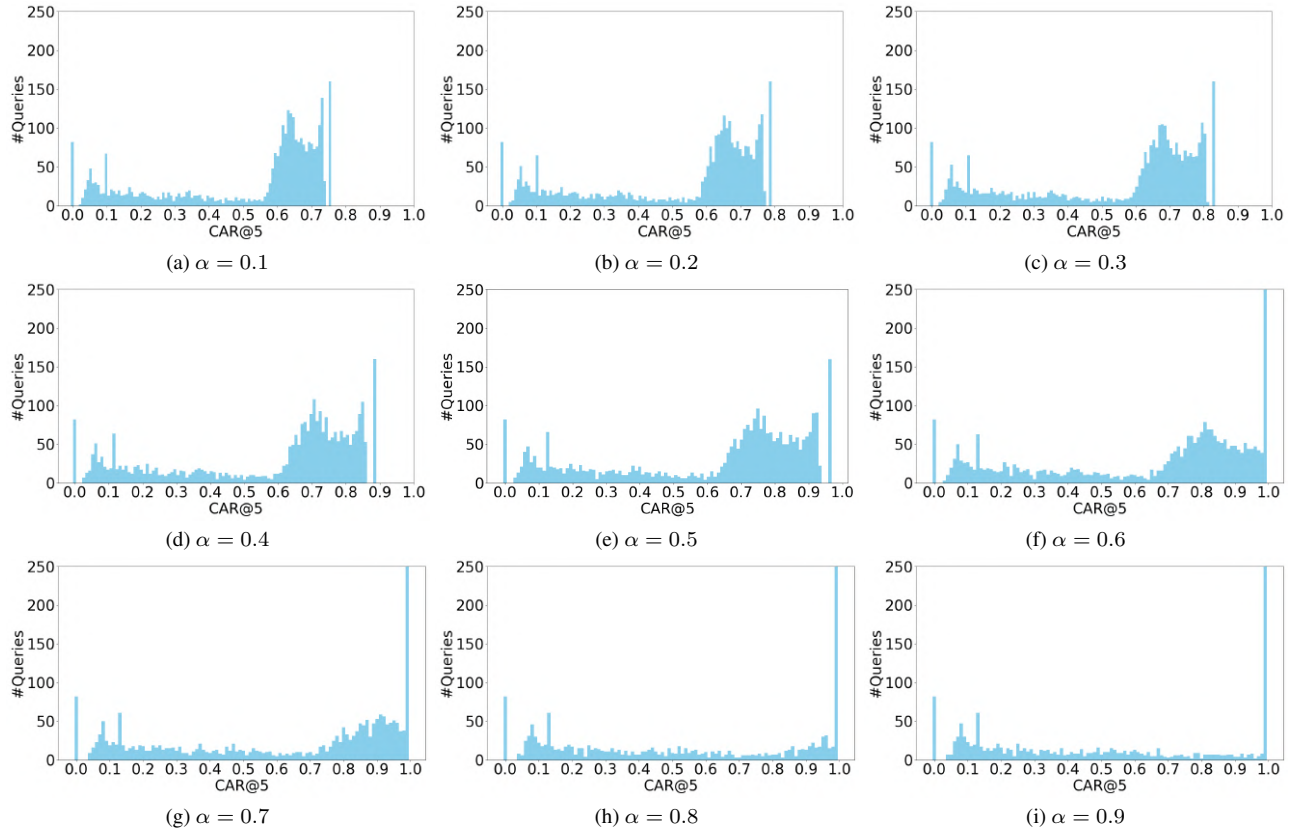
Figure 10. Statistical overview of SciGA-145k across top-level arXiv categories. (a) Average token length of titles, (b) average token length of abstracts, (c) average token length of full-texts, (d) average token length of captions, (e) number of figures per paper, and (f) number of subfigures per paper. Each graph presents the mean and standard deviation for each top-level arXiv category, alongside the overall micro-average. These statistics highlight category-specific variations and overall distribution trends in the dataset.

timates even for flat or ambiguous predictions, inflating CAR@k and reducing its discriminative power. We adopt $\alpha = 0.5$ as a balanced setting, offering interpretable and discriminative scores. Notably, Changing α does not alter query rankings, ensuring CAR remains valid for comparative evaluation.

D. Equipment Details

This section outlines the experimental setup used in our study, including backbone models, hyper-parameter settings, and computational resources.

Model Architectures. Our experiments utilized representative pretrained backbone models listed in Tab. 5. Among the tested methods, (iv) Abs2Fig w/cap involved a lightweight architectural modification that integrates figure and caption embeddings via the Hadamard product. The overall structure



(a) $\alpha = 0.1$ (b) $\alpha = 0.2$ (c) $\alpha = 0.3$

(d) $\alpha = 0.4$ (e) $\alpha = 0.5$ (f) $\alpha = 0.6$

(g) $\alpha = 0.7$ (h) $\alpha = 0.8$ (i) $\alpha = 0.9$

Figure 11. Distribution of CAR@5 scores across test queries for different values of α. Each histogram represents the number of queries (#Queries) for a given CAR@5 score, illustrating how the score distribution shifts as α increases. At lower α values, CAR@5 scores are more compressed, while higher α values lead to a broader spread with an increasing concentration near 1.0.

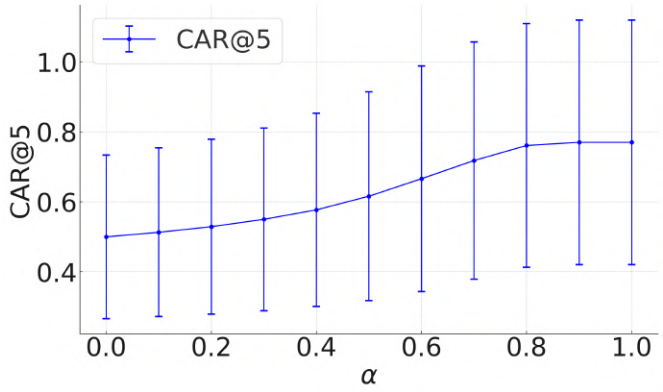
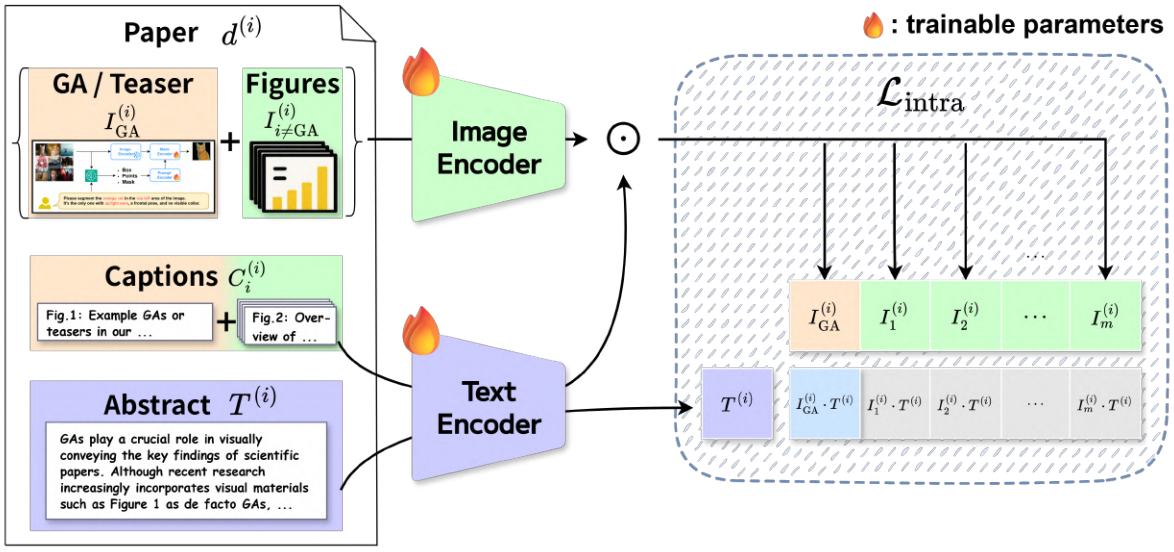


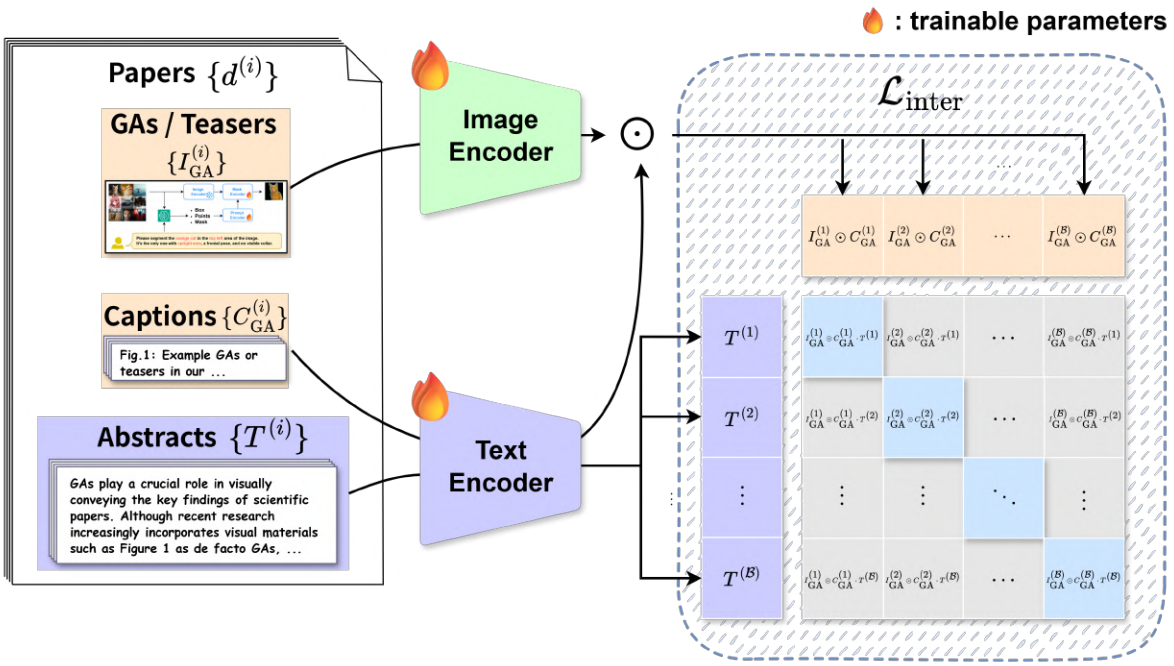
Figure 12. Mean and standard deviation of CAR@5 scores across test queries for different values of α. As α increases, the average CAR@5 score gradually rises, indicating reduced penalization effects on model's confidence.

is illustrated in Fig. 13.

Hyper-parameter Settings. We summarize the final hyper-parameter settings for all models in Tab. 6. All models were trained and evaluated under a hold-out validation protocol, with learning rate, batch size, and the number of sampled figures per paper m in Intra-GA Recommendation explored over $\{1 \times 10^{-7}, 5 \times 10^{-7}, 1 \times 10^{-6}, 5 \times 10^{-6}, 1 \times 10^{-5}\}$, $\{64, 128, 256, 512, 1024, 2048\}$, and $\{5, 6, 7, 8, 9\}$, respectively. Other hyper-parameters followed prior literature or official implementations without further tuning. All reported scores are averaged over five independent runs with different random seeds to account for training variability.



(a) Intra-GA Recommendation



(b) Inter-GA Recommendation

Figure 13. Overview of the contrastive learning framework for method (iv) Abs2Fig w/cap applied to (a) Intra-GA Recommendation and (b) Inter-GA Recommendation. Both frameworks encode figures and texts (abstracts and captions) separately into embeddings, optimizing contrastive losses ($\mathcal{L}_{\text{Intra}}$, $\mathcal{L}_{\text{Inter}}$) to align semantically or visually related pairs. The flame icon indicates trainable model components.

Computational Environment. Experiments were conducted using an NVIDIA RTX A6000 GPU (48 GB). Training and evaluation took approximately 12 hours for Intra-GA Recommendation and 8 hours for Inter-GA Recommendation.

E. Additional Results

E.1. Intra-GA Recommendation

Qualitative Evaluation. We analyze representative cases from the best-performing model (Long-CLIP [61] within

Table 5. Pretrained Weights for Backbone Models.

Method	Backbone Model	Pre-trained Weight
(i) Abs2Cap	BERTScore [62]	`allenai/scibert_scivocab_uncased`
(ii) GA-binCl	EfficientNetV2 [47]	`EfficientNet_V2_L_Weights.IMAGENET1K_V1`
	ViT [9]	`google/vit-large-patch16-224-in21k`
	CLIP image encoder [39]	`openai/clip-vit-large-patch14`
	SwinTransformerV2 [29]	`microsoft/swin-large-patch4-window7-224-in22k`
	ConvNeXtV2 [55]	`facebook/convnextv2-large-22k-224`
(iii) Abs2Fig (iv) Abs2Fig w/cap	CLIP [39]	`openai/clip-vit-large-patch14`
	BLIP-2 [26]	`Salesforce/blip2-itm-vit-g`
	X^2-VLM [60]	`X2VLM-large (4M)`
	OpenCLIP [7]	`laion/CLIP-ViT-L-14-laion2B-s32B-b82K`
	SigLIP2 [50]	`google/siglip2-large-patch16-256`
	Long-CLIP [61]	`BeichenZhang/LongCLIP-L`

Table 6. Detailed training hyper-parameters.

Hyper-Parameter	Value
For (ii) GA-binCl	
Epochs	30
Batch Size	128
Learning Rate	5e-7
LR Scheduler	Cosine Annealing
Mixed Precision Training	Enabled (AMP)
Weight Decay	1e-3
AdamW β_1	0.9
AdamW β_2	0.999
AdamW ϵ	1e-8
Loss Weighting	Inverse Frequency
For (iii) Abs2Fig / (iv) Abs2Fig w/cap	
Epochs	15
Batch Size	1024
Sampled Figures per Paper m	7
Learning Rate	1e-6
LR Scheduler	Cosine Annealing
Mixed Precision Training	Enabled (AMP)
Weight Decay	1e-3
AdamW β_1	0.9
AdamW β_2	0.999
AdamW ϵ	1e-8
Temperature τ	0.07

method (iv) Abs2Fig w/cap) to understand its behavior in Intra-GA Recommendation. Figs. 14, 15, 16 and 17 illustrate typical outputs, and corresponding CAR and nDCG [5]. The model tends to favor figures with architectural overviews or grid-style layouts, which frequently preferred as GAs or teasers by authors. When multiple candidates share such designs, scores converge and CAR appropriately reflects this ambiguity with moderate values. This nuance is often missed by conventional metrics. In contrast, charts are generally scored lower, suggesting a learned preference for structured, concept-focused visuals. These results indicate that the model captures not only text-figure relevance but also topic-specific visual conventions linked to effective GA design. CAR quantifies this instance-level behavior, unifying relevance and confidence into a single interpretable score.

Distributional Analysis of CAR. Fig. 18 shows CAR@5 distributions for each method. Abs2Cap displays a polarized pattern, working well only with strong lexical cues. GA-binCl yields more moderate scores (0.6–0.8) but lacks high-performing cases. Abs2Fig achieves stronger performance overall (0.7–0.9), and adding captions (Abs2Fig w/-cap) shifts scores above 0.9 with fewer failures. These results show that combining captions with visual features improves recommendation quality and consistency.

Domain-specific and Cross-domain Analysis. This experiment evaluates how well our GA recommendation model generalizes across scientific domains. We trained the Abs2Fig w/cap model using Long-CLIP independently on three domains: *math*, *cond-mat*, and *astro-ph*. For each domain, the model was trained, validated, and tested on papers from that domain only, using the same retrieval setting as in the main paper: given an abstract, rank all candidate figures from the same paper and compute R@1, DCG@5, and CAR@5. Tab. 7 shows in-domain performance. The *cond-mat* model achieves the highest scores across all metrics, while *math* and *astro-ph* perform notably worse, suggesting that GA identification is harder in these fields, possibly due to smaller datasets or less standardized figure styles. We then evaluated each model on the test sets of the other domains without fine-tuning to assess cross-domain transferability. Fig. 19 summarizes the results. Models trained on the `cs` domain generalize best to all other domains, likely because of the larger training scale and more consistent GA formats (e.g., model diagrams, algorithmic flows). In contrast, mod-

els trained on *math* or *astro-ph* transfer poorly, reflecting narrower visual semantics and limited training data. These findings demonstrate that domain coverage and visual diversity in training data strongly affect GA recommendation performance and that cross-domain modeling or adaptive approaches could further improve generalization.

E.2. Inter-GA Recommendation

Qualitative Evaluation. Fig. 20 presents examples of Inter-GA Recommendation results obtained using different methods. We observe that methods based on CLIP-like models, such as Abs2Fig and Abs2Fig w/cap, are capable of retrieving GAs from papers that share similar topics with the query. Notably, this is achieved even though the retrieval process relies solely on the abstract of the query and the GAs of candidate papers, without explicitly incorporating the abstracts or full-texts of the recommended papers into the search context. While these methods effectively capture topic-level alignment, they may lack the capacity to suggest GAs that introduce surprising or serendipitous ideas from different research areas. Enabling such cross-domain inspiration may require additional mechanisms to intentionally diversify the recommendation results beyond semantic similarity.

Quantitative Analysis. Tab. 8 summarizes the top-k mean and standard deviation along the four axes introduced in Sec. 4.1.2. CLIP-based models exhibit consistently higher Semantic and Visual Coherence, indicating that contrastive encoders effectively retrieve GAs that are both topically and visually aligned with the query. Long-CLIP achieves the strongest Visual Coherence on average, while caption-based lexical methods (Abs2Cap) remain weak across all axes. The variance patterns further distinguish model behaviors: random sampling naturally yields the largest diversity, whereas contrastive models produce tighter, more coherent top-k sets. These observations confirm that the four axes provide complementary perspectives for characterizing Inter-GA retrieval quality.

Robustness Check with DreamSim. To evaluate the robustness of the Visual Coherence axis, we additionally computed nDCG@k and top-k mean similarity using DreamSim [10], an alternative perceptual similarity metric. As shown in Tab. 9, results under DreamSim exhibit trends consistent with those obtained from CLIPScore [13], indicating that the visual consistency of recommended GAs is similarly captured across different perceptual similarity measures.

F. User Study Interface and Protocol

We present the format and participant demographics of our user study, which was conducted via an online questionnaire (Google Form). Fig. 21 shows the questionnaire format and questions. The study involved 15 participants: 3 master's students, 6 industry researchers with master's degrees, 2 Ph.D. students, and 4 Ph.D. holders. All participants had prior experience designing GAs or teasers and peer-reviewed publication.

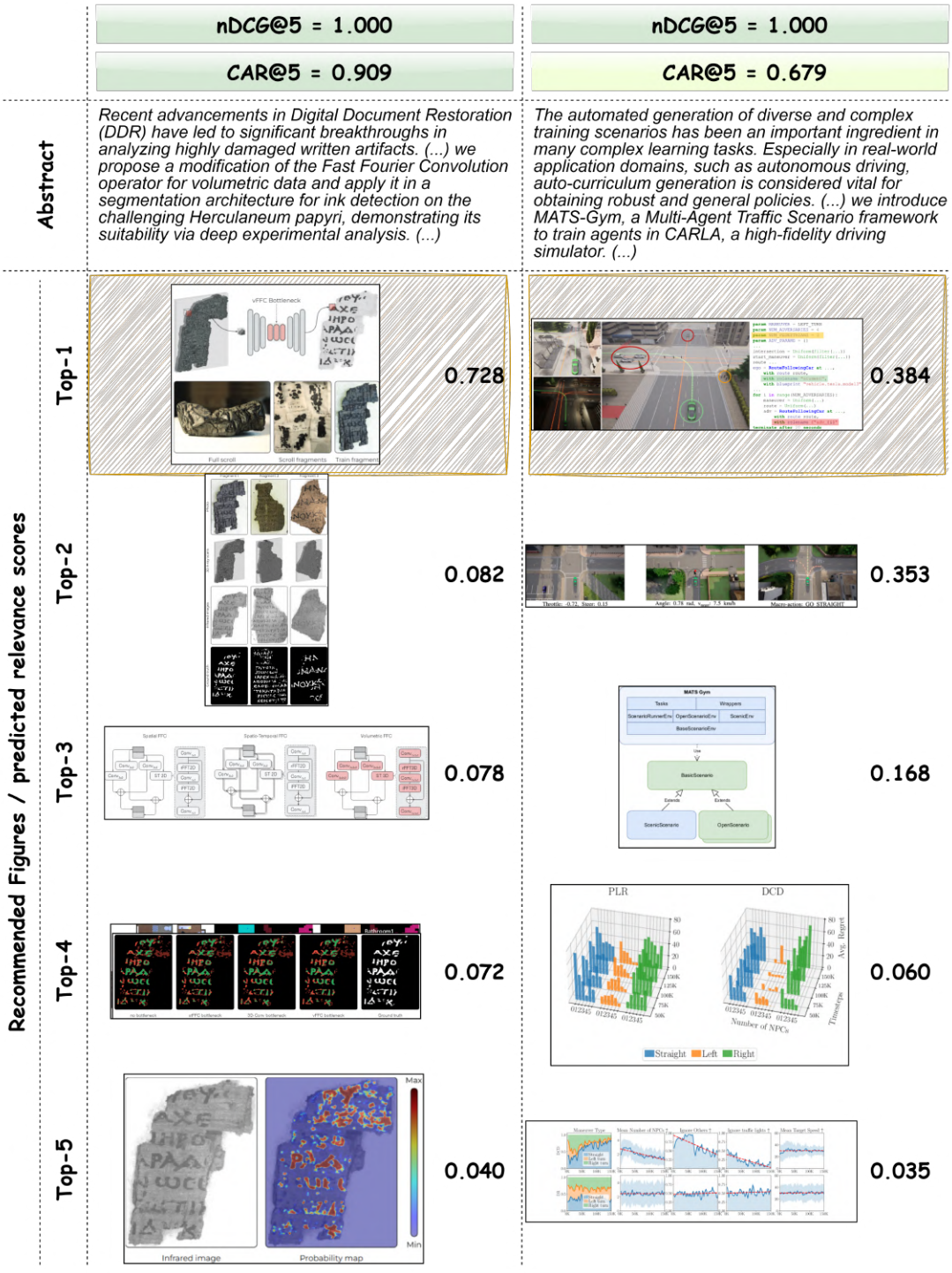


Figure 14. Qualitative examples of Intra-GA Recommendation results obtained by the best-performing model (Long-CLIP within method (iv) Abs2Fig w/cap). [9] The yellow-highlighted figures represent GTs.

[9] arXiv ID: 2308.05070, 2403.17805

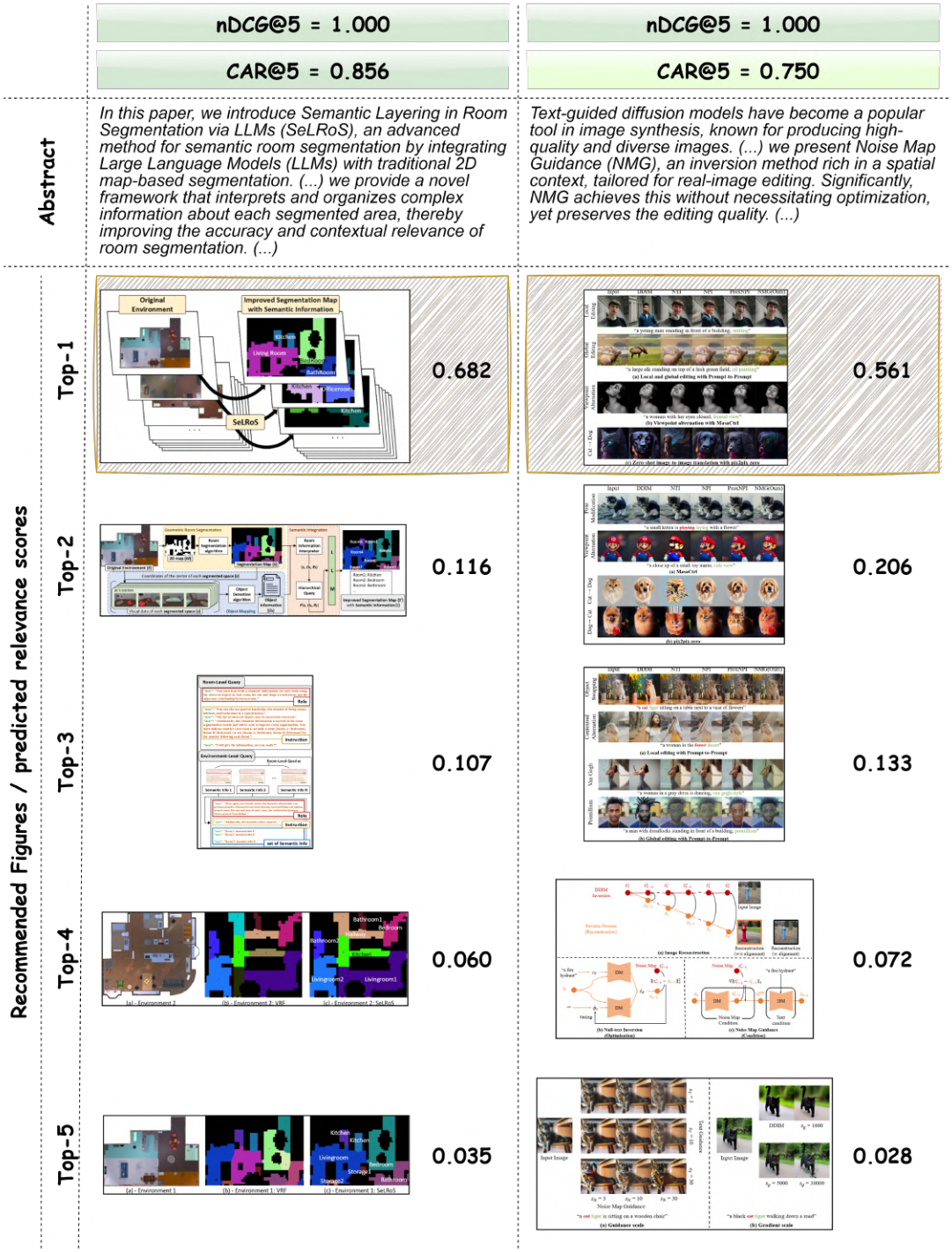


Figure 15. Qualitative examples of Intra-GA Recommendation results obtained by the best-performing model (Long-CLIP within method (iv) Abs2Fig w/cap). [10] The yellow-highlighted figures represent GTs.

[10] arXiv ID: 2403.12920, 2402.04625

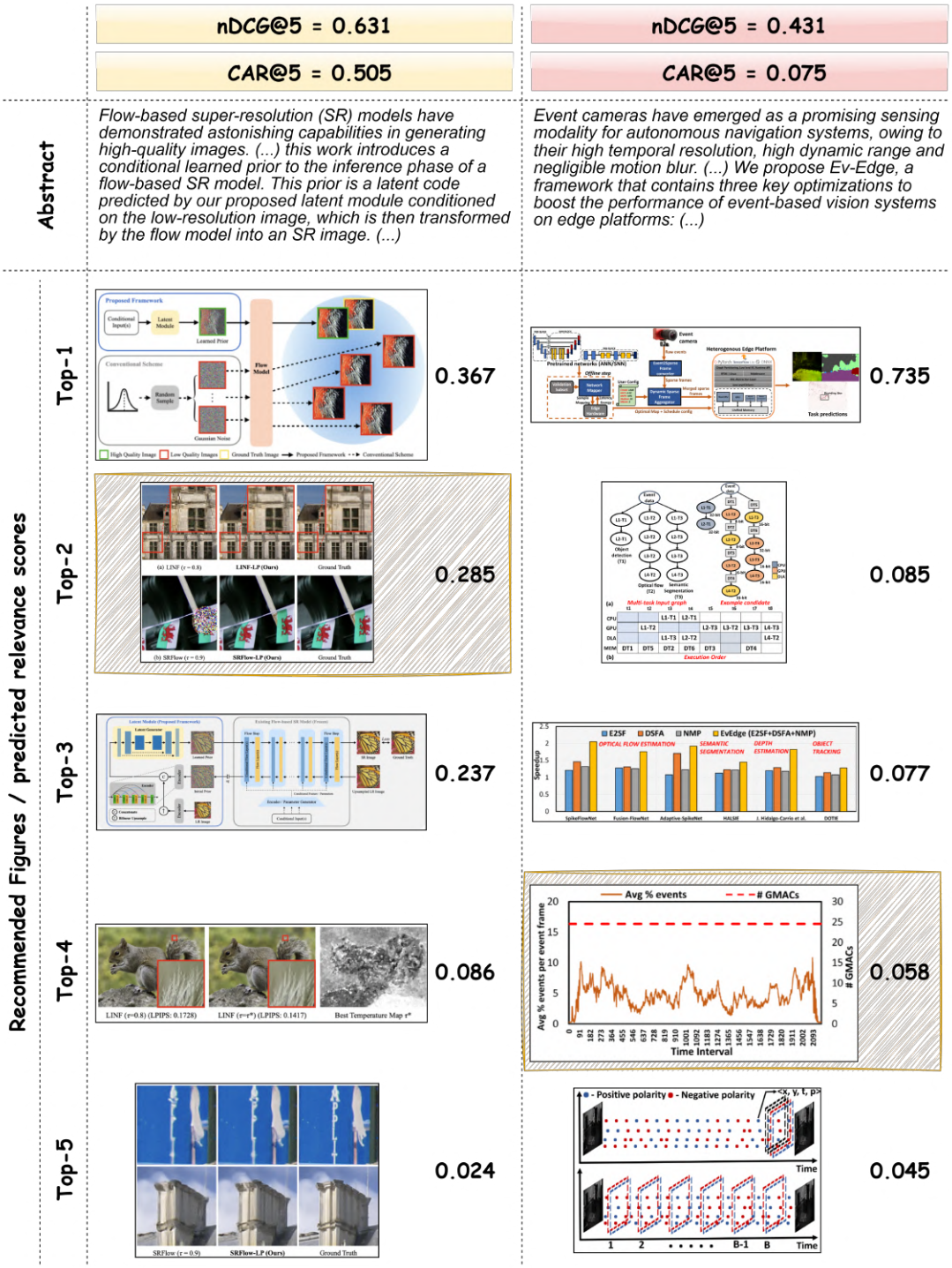


Figure 16. Qualitative examples of Intra-GA Recommendation results obtained by the best-performing model (Long-CLIP within method (iv) Abs2Fig w/cap). [11] The yellow-highlighted figures represent GTs.

[11] arXiv ID: 2403.10988, 2403.15717

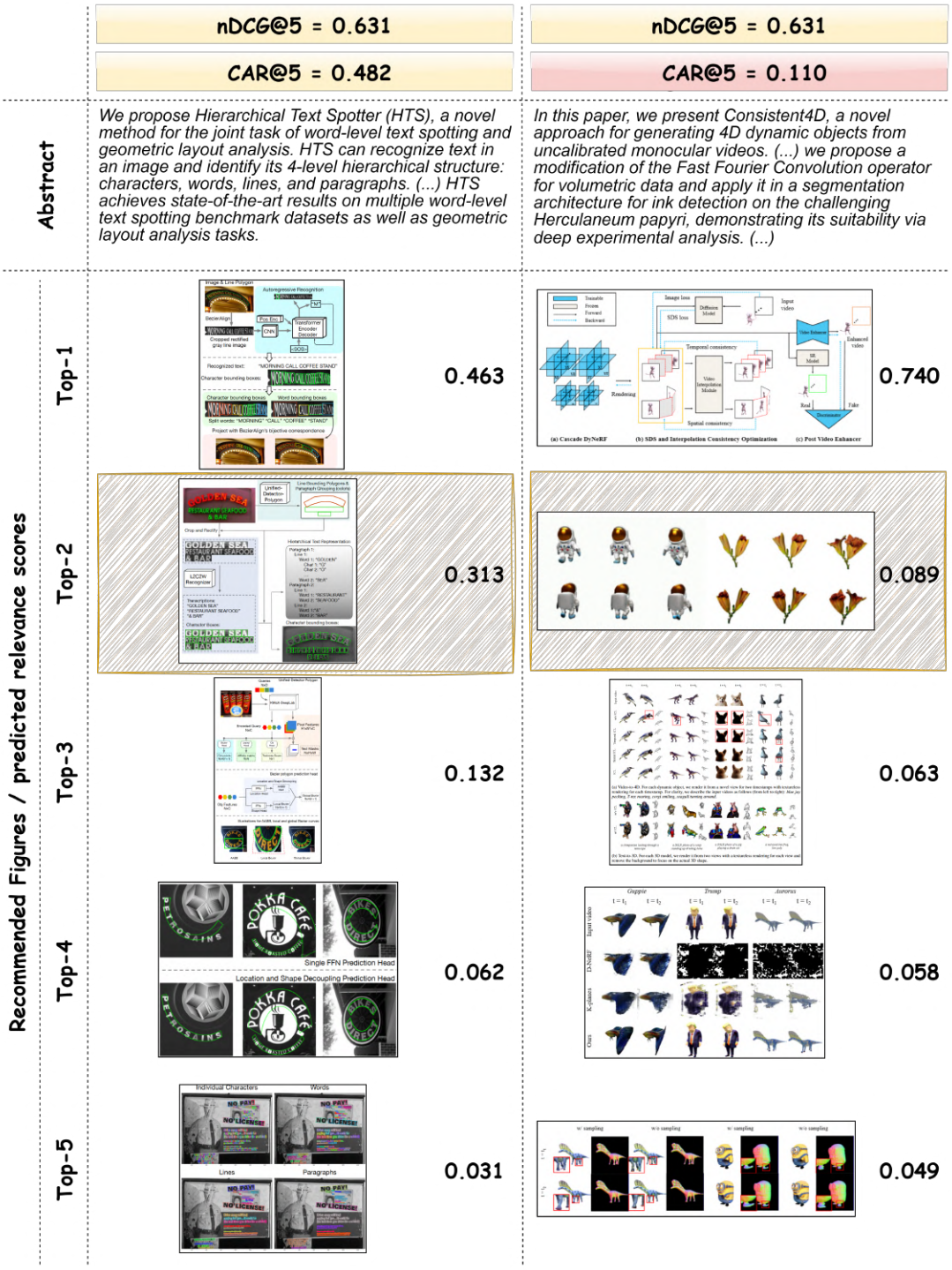
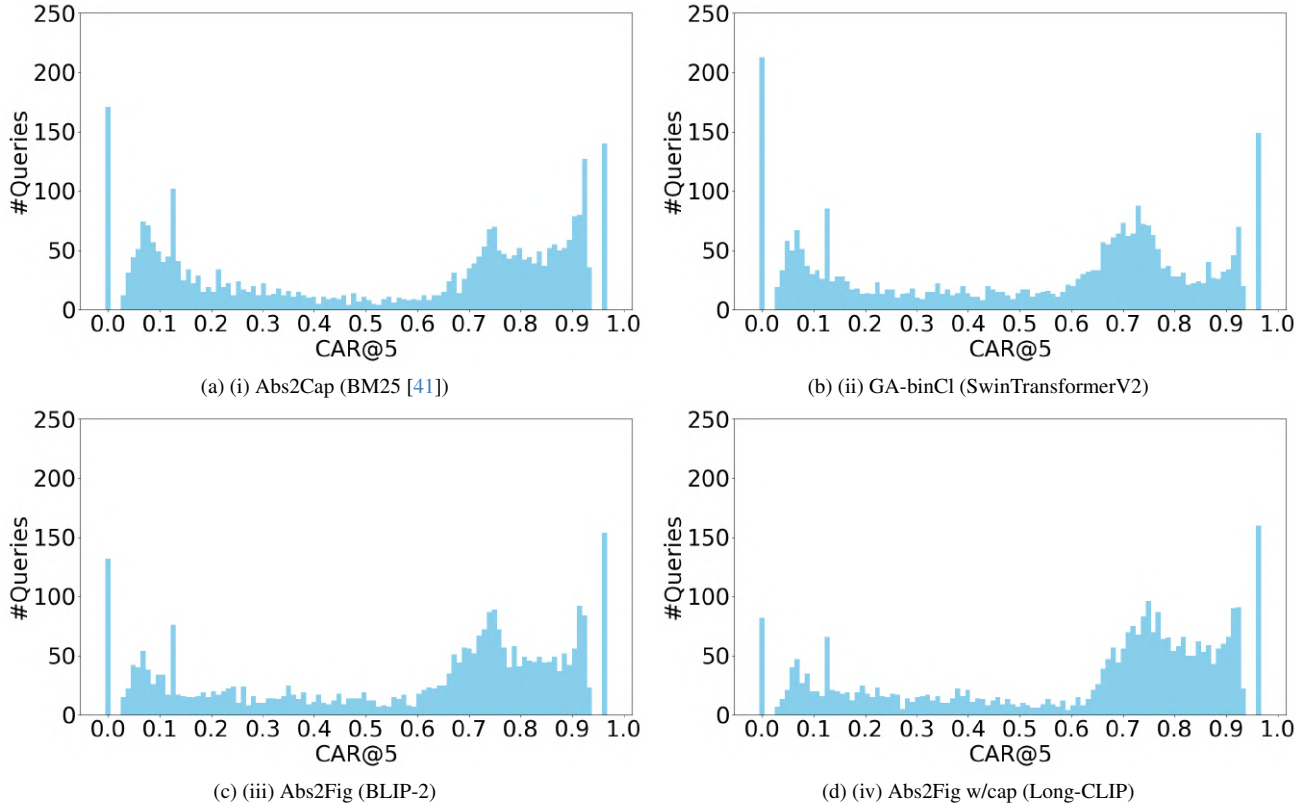


Figure 17. Qualitative examples of Intra-GA Recommendation results obtained by the best-performing model (Long-CLIP within method (iv) Abs2Fig w/cap). [12] The yellow-highlighted figures represent GTs.

[12] arXiv ID: 2310.17674, 2311.02848



(a) (i) Abs2Cap (BM25 [41])

(b) (ii) GA-binCl (SwinTransformerV2)

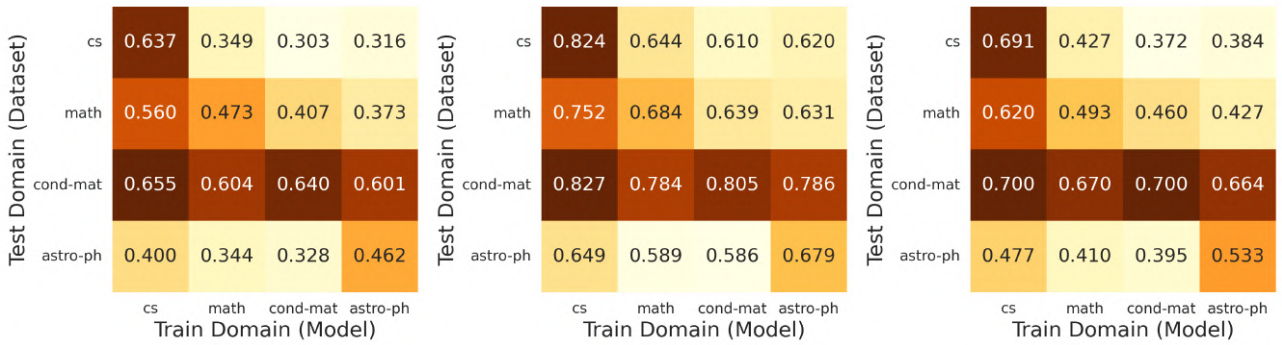
(c) (iii) Abs2Fig (BLIP-2)

(d) (iv) Abs2Fig w/cap (Long-CLIP)

Figure 18. Distribution of CAR@5 scores across individual queries for the best-performing models in each Intra-GA Recommendation method. Higher CAR@5 values indicate higher model's confidence and more reliable top-ranked GA recommendations. Methods with distributions skewed toward higher values reflect stronger model confidence and more effective recommendation performance.

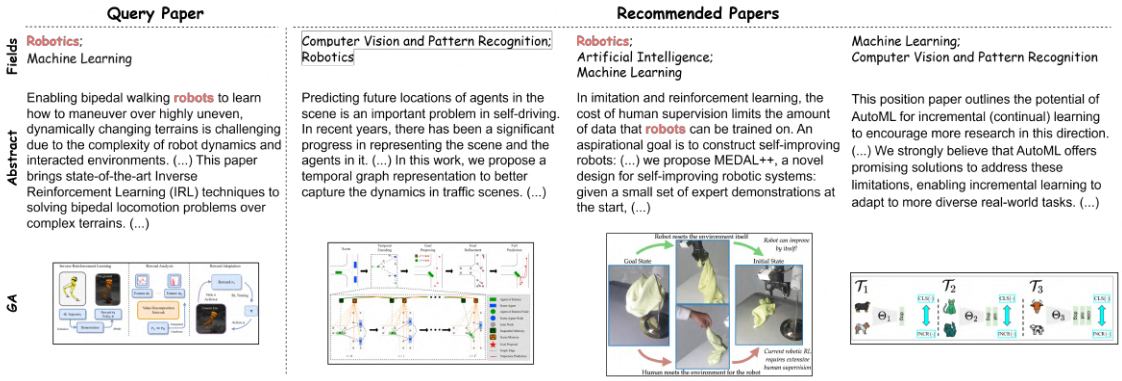
Table 7. In-domain performance of Intra-GA Recommendation models across scientific fields. 0.5 ↑ indicates the proportion of semantically justifiable predictions (CAR@5 > 0.5). The `cond-mat` domain shows the highest CAR@5 values, suggesting that GA selection is more consistent and easier to model in this field.

Research Fields	**Data Size**	**R@1**	**R@2**	**R@3**	**MRR**	**nDCG@5**	**CAR@5**	
							Mean	0.5 ↑
cs	20,520	0.637	0.826	0.914	0.778	0.824	0.615	0.691
math	1,498	0.473	0.663	0.763	0.643	0.684	0.493	0.493
cond-mat	3,323	0.640	0.805	0.871	0.767	0.805	0.639	0.700
astro-ph	1,949	0.462	0.651	0.764	0.633	0.679	0.494	0.533

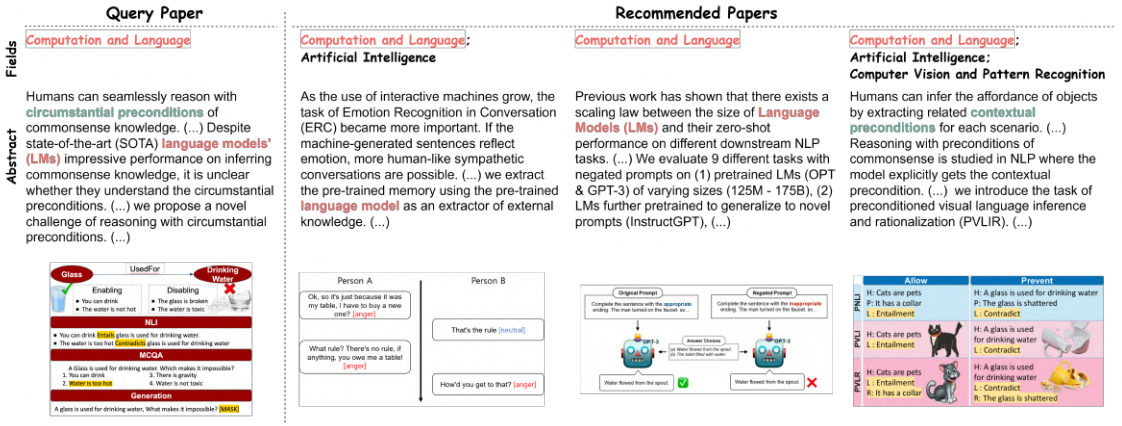


(a) R@1 (b) nDCG@5 (c) CAR@5: 0.5↑

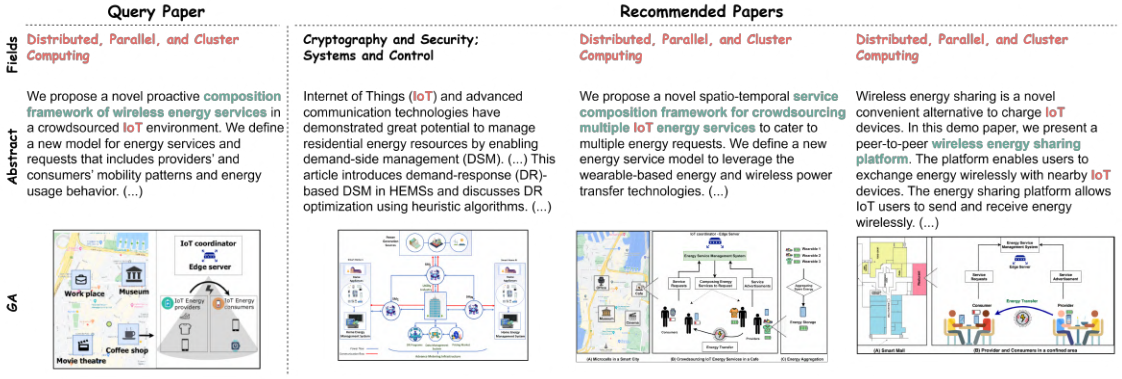
Figure 19. Cross-domain evaluation of Intra-GA Recommendation models trained on different scientific fields. We report R@1, nDCG@5, and proportion of predictions with CAR@5 exceeding 0.5. Models trained on any domain performed well on the `cond-mat` test set, indicating that GAs in this field are relatively easy to identify. `cs`-trained models generalize well, possibly due to the widespread use of teaser figures in that field.



(a) (i) Abs2Cap (ROUGE-L) [13]



(b) (iii) Abs2Fig (CLIP) [14]



(c) (iv) Abs2Fig w/cap (CLIP) [15]

Figure 20. Examples of Inter-GA Recommendation results obtained by different methods. Pink-highlighted research fields or keywords within abstracts indicate matching primary research categories. Green-highlighted phrases denote topic-level relevance. These results highlight the different characteristics of the recommendation methods. (a) Abs2Cap (ROUGE-L [8]) produces diverse recommendations, retrieving papers from a broad range of topics. In contrast, (b) Abs2Fig (CLIP) and (c) Abs2Fig w/cap (CLIP) focus on recommending GAs from papers that share similar topics with the query paper, emphasizing strong semantic alignment within the same research domain.

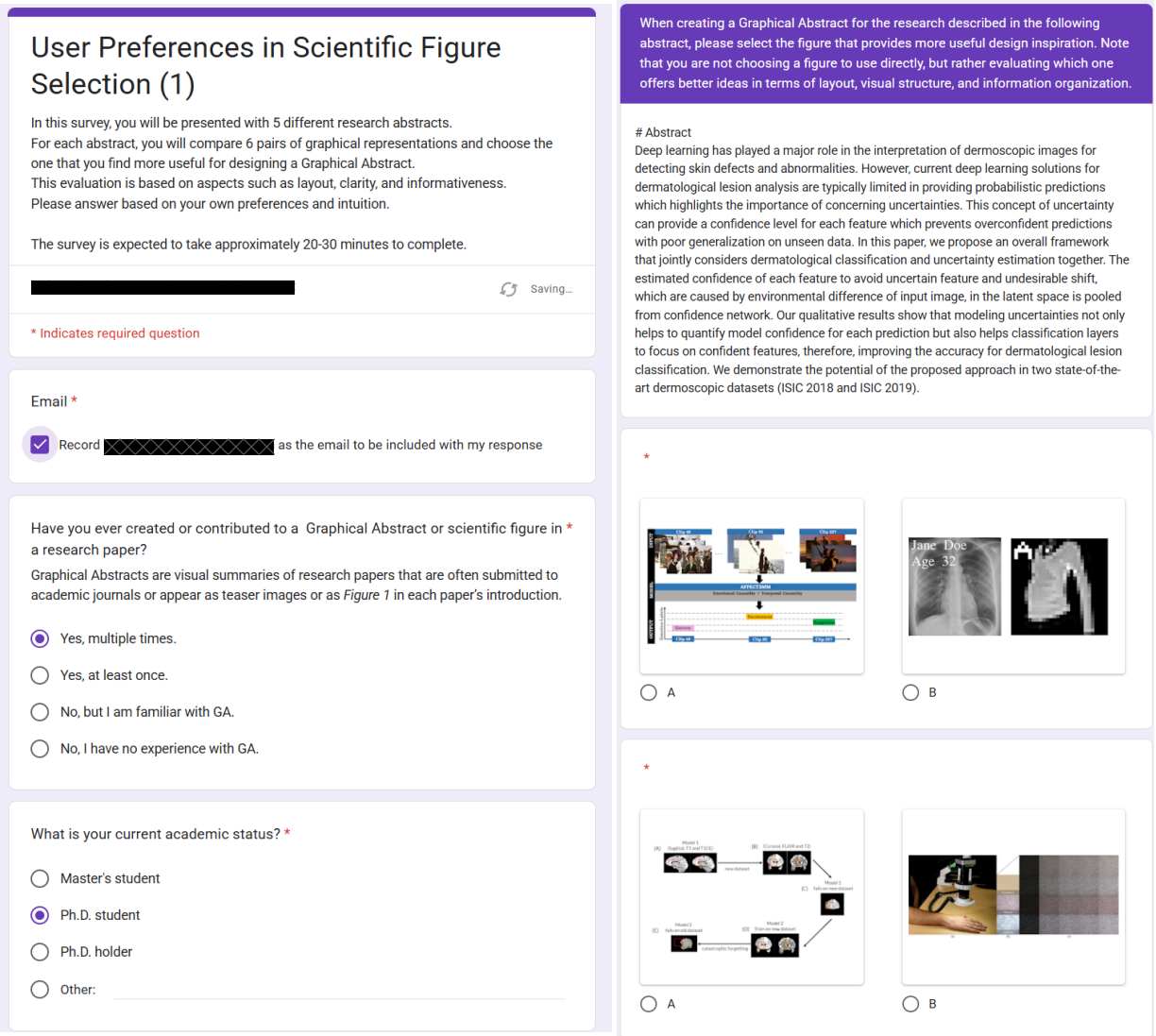
[13](a) arXiv ID (Query): 2309.16074, arXiv ID (Recommended): 2207.00255, 2303.01488, 2311.11963
[14](b) arXiv ID (Query): 2104.08712, arXiv ID (Recommended): 2108.11626, 2209.12711, 2306.01753
[15](c) arXiv ID (Query): 2107.12519, arXiv ID (Recommended): 2109.11627, 2308.09886, 2208.13506

Table 8. Quantitative comparison of various approaches for Inter-GA Recommendation. For each method, we report the mean and standard deviation of its top-k recommended GAs under four axes introduced in Sec. 4.1.2: Field Match, Semantic Coherence, Visual Coherence, and Aesthetic Quality. Higher values indicate stronger alignment along each respective axis, and larger standard deviations reflect greater diversity within the top-k set. Best scores for each metric are shown in **bold**, and the highest standard deviations are underlined.

Method	Backbone	Field Match@k		Semantic Coherence@k		Visual Coherence@k		Aesthetics Quality@k	
		top-5	top-10	top-5	top-10	top-5	top-10	top-5	top-10
(BL) Random Sampling	-	0.338	0.345	0.227 ± 0.111	0.228 ± 0.115	$\underline{0.545 \pm 0.077}$	$\underline{0.545 \pm 0.081}$	$\underline{0.125 \pm 0.248}$	$\underline{0.125 \pm 0.253}$
(i) Abs2Cap	ROUGE-L [8]	0.502	0.486	0.314 ± 0.114	0.306 ± 0.118	0.579 ± 0.066	0.578 ± 0.069	0.136 ± 0.236	0.136 ± 0.243
	METEOR [2]	0.421	0.417	0.268 ± 0.110	0.264 ± 0.112	0.573 ± 0.063	0.571 ± 0.064	0.130 ± 0.240	0.130 ± 0.249
	CIDEr [53]	0.438	0.420	0.287 ± 0.105	0.273 ± 0.108	0.579 ± 0.064	0.577 ± 0.066	0.134 ± 0.237	0.135 ± 0.243
	BM25 [41]	0.704	0.685	0.489 ± 0.105	0.468 ± 0.111	0.605 ± 0.072	0.601 ± 0.074	0.151 ± 0.233	0.148 ± 0.240
	BERTScore [62]	0.549	0.545	0.360 ± 0.107	0.351 ± 0.109	0.580 ± 0.069	0.578 ± 0.071	0.147 ± 0.232	0.145 ± 0.239
(iii) Abs2Fig	CLIP [39]	0.729	0.719	0.455 ± 0.105	0.444 ± 0.109	0.646 ± 0.054	0.642 ± 0.057	$\mathbf{0.160 \pm 0.220}$	$\mathbf{0.162 \pm 0.227}$
	BLIP-2 [26]	0.683	0.674	0.419 ± 0.110	0.410 ± 0.114	0.622 ± 0.063	0.620 ± 0.065	0.152 ± 0.228	0.153 ± 0.235
	X²-VLM [60]	0.418	0.402	$\underline{0.263 \pm 0.116}$	$\underline{0.257 \pm 0.122}$	0.461 ± 0.032	0.451 ± 0.033	0.133 ± 0.206	0.122 ± 0.203
	OpenCLIP [7]	0.720	0.710	0.451 ± 0.106	0.440 ± 0.109	0.632 ± 0.058	0.630 ± 0.061	0.159 ± 0.221	0.157 ± 0.229
	SigLIP2 [50]	0.631	0.620	0.387 ± 0.110	0.381 ± 0.114	0.598 ± 0.065	0.597 ± 0.068	0.142 ± 0.229	0.144 ± 0.239
	Long-CLIP [61]	0.726	0.717	0.456 ± 0.108	0.445 ± 0.103	$\mathbf{0.648 \pm 0.056}$	$\mathbf{0.644 \pm 0.060}$	0.159 ± 0.221	0.160 ± 0.228
(iv) Abs2Fig w/cap	CLIP [39]	**0.755**	**0.742**	0.493 ± 0.098	0.479 ± 0.101	0.614 ± 0.067	0.611 ± 0.071	0.152 ± 0.229	0.152 ± 0.237
	BLIP-2 [26]	0.647	0.639	0.390 ± 0.105	0.382 ± 0.109	0.597 ± 0.067	0.596 ± 0.068	0.147 ± 0.228	0.149 ± 0.236
	X²-VLM [60]	0.415	0.399	0.254 ± 0.114	0.250 ± 0.119	0.555 ± 0.067	0.552 ± 0.072	0.143 ± 0.213	0.140 ± 0.226
	OpenCLIP [7]	0.749	0.737	0.489 ± 0.097	0.475 ± 0.100	0.615 ± 0.066	0.611 ± 0.069	0.154 ± 0.231	0.150 ± 0.237
	SigLIP2 [50]	0.235	0.336	0.186 ± 0.128	0.212 ± 0.129	0.462 ± 0.076	0.467 ± 0.079	0.134 ± 0.253	0.117 ± 0.247
	Long-CLIP [61]	0.753	0.737	$\mathbf{0.498 \pm 0.098}$	$\mathbf{0.482 \pm 0.103}$	0.614 ± 0.070	0.611 ± 0.073	0.148 ± 0.231	0.145 ± 0.240

Table 9. Additional quantitative results for Inter-GA Recommendation using DreamSim as an alternative perceptual similarity metric. We report (1) nDCG@k using DreamSim-based pseudo-relevance labels, and (2) the mean and standard deviation of DreamSim similarities over the top-k retrieved GAs. These results examine robustness to the choice of visual similarity metric. Best values are shown in **bold**, and the highest standard deviations are underlined.

Method	Backbone	nDCG@k (DreamSim)			DreamSim@k	
		top-5	top-10	top-30	top-5	top-10
(BL) Random Sampling	-	0.679	0.711	0.803	$\underline{0.382 \pm 0.086}$	0.382 ± 0.090
(i) Abs2Cap	ROUGE-L [8]	0.722	0.750	0.830	0.417 ± 0.082	0.415 ± 0.086
	METEOR [2]	0.749	0.777	0.851	0.433 ± 0.067	0.431 ± 0.071
	CIDEr [53]	0.742	0.771	0.847	0.425 ± 0.072	0.425 ± 0.075
	BM25 [41]	0.742	0.767	0.841	0.443 ± 0.081	0.438 ± 0.083
	BERTScore [62]	0.704	0.735	0.821	0.406 ± 0.083	0.406 ± 0.086
(iii) Abs2Fig	CLIP [39]	**0.753**	**0.778**	**0.853**	0.456 ± 0.072	0.452 ± 0.075
	BLIP-2 [26]	0.752	**0.778**	**0.853**	0.456 ± 0.070	0.453 ± 0.073
	X²-VLM [60]	0.438	0.437	0.652	0.245 ± 0.036	0.232 ± 0.037
	OpenCLIP [7]	0.748	0.776	0.851	0.453 ± 0.070	0.451 ± 0.073
	SigLIP2 [50]	0.726	0.756	0.838	0.432 ± 0.073	0.431 ± 0.076
	Long-CLIP [61]	0.752	**0.778**	0.852	$\mathbf{0.457 \pm 0.072}$	$\mathbf{0.454 \pm 0.075}$
(iv) Abs2Fig w/cap	CLIP [39]	0.739	0.766	0.843	0.447 ± 0.077	0.443 ± 0.080
	BLIP-2 [26]	0.764	0.789	0.860	0.453 ± 0.065	0.450 ± 0.069
	X²-VLM [60]	0.697	0.726	0.816	0.387 ± 0.074	0.385 ± 0.078
	OpenCLIP [7]	0.744	0.772	0.847	0.449 ± 0.074	0.446 ± 0.077
	SigLIP2 [50]	0.460	0.513	0.680	0.255 ± 0.073	$\underline{0.270 \pm 0.103}$
	Long-CLIP [61]	0.738	0.765	0.841	0.447 ± 0.079	0.443 ± 0.081



(a) (b)

Figure 21. Screenshot of the questionnaire used in the user study. (a) The introductory section of the questionnaire, asking participants about their prior experience with GAs and their current academic status. (b) Example of the comparative evaluation task. After reading an abstract, participants were presented with pairs of figures recommended by different methods and asked to select the one they found more useful as a design reference when creating a new GA.