Beyond Single Images: A Comprehensive Benchmark for Album-Level Vision-Language Understanding

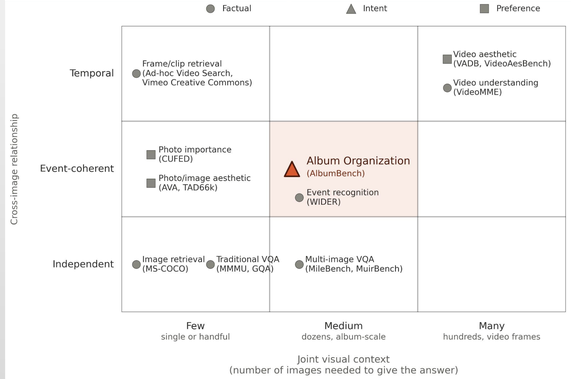
Shawn Huang*; Brian Price†; Yifei Fan†; Bryan Morse*
*Brigham Young University, †Adobe Research

CVPR JUNE 3-7, 2026 DENVER COLORADO

Motivation

- 1.9 trillion photos were taken worldwide in 2024[1]
- Finding, sorting, and curating photos from albums is tedious
- Can we automate this album organization process?
- No existing benchmarks/datasets address album organization

Album Organization



What is unique about album organization:

- it is based on **user intent**;
- it encompasses **multiple tasks**;
- it is **not image retrieval**, users look back and forth to compare;
- it is **not VQA**, because the answers are not based on the facts about the images but what the user intended;
- its **image distribution is different** than single images or video frames: highly related but visually distinct
- it requires **contextual understanding** of the entire image pool (album)

We contribute AlbumBench:

- Built upon CUFED[2] (23 event types: personal event to holiday)
- 641 albums; 27,051 images; 5 annotations per image
- Train/test: 80/20
- Hold-out for testing: *casual_family_gathering* and *beach_trip*

AlbumBench Tasks

Intent Selection:

*Query: Capture the essence of traditional wedding customs and **rituals** for a cultural study paper

Intent Rating:

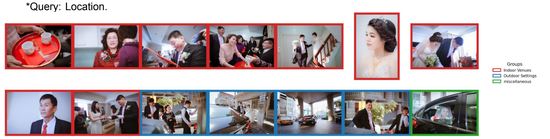
*Query: Gather photographs that depict joyful expressions and celebrations to create a wedding highlight video

Group Labeling:

*Query: Event Phase. Available groups: ["Preparation", "Ceremony"]

Group Clustering:

*Query: Location.



Benchmarking Methods

Models:

- Proprietary:
  - GPT-5
  - Gemini-2.5-Pro
- Open-source:
  - InternVL-3.5 (8B, 38B)
  - Qwen3-VL (8B, 32B, 235B-A22B)
  - Keye-VL-1.5 (8B)
- Caption Baseline:
  - Gemini-Caption-Short
  - Gemini-Caption-Long

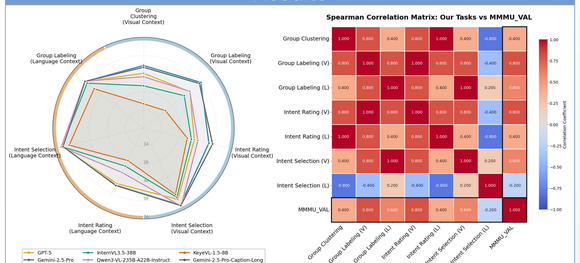
Instruct and Thinking:

- For each config on the left, we include **Instruct** and **Thinking** variants
- Results in **20** model configurations

Contextual info:

- visual context: gives models the actual images
- language context: replace each album with its caption

Results



Model	Intent Selection F1↑	Prec.↑	Recall↑	mAP↑	Intent Rating Acc.↑	MAE↓	RMSE↓	Group Labeling ARI↑	F1↑	Jaccard↑	NMI↑	Group Clustering ARI↑	F1↑	Jaccard↑	NMI↑
Open-source Instruct VLMs															
Keye-VL-1.5-8B	0.601	0.583	0.778	0.549	0.367	0.876	1.133	0.23	0.47	0.371	0.31	0.199	0.37	0.286	0.305
InternVL3.5-8B	0.416	0.636	0.374	0.296	0.355	0.904	1.173	0.24	0.481	0.377	0.338	0.205	0.404	0.307	0.327
Qwen3-VL-8B	0.591	0.538	0.804	0.577	0.4	0.816	1.08	0.205	0.505	0.401	0.272	0.222	0.416	0.325	0.313
Qwen3-VL-32B	0.682	0.621	**0.868**	**0.647**	**0.46**	0.718	0.984	0.424	0.635	0.537	0.483	0.376	0.471	0.384	0.476
InternVL3.5-38B	0.629	0.609	0.769	0.586	0.402	0.797	1.044	0.426	0.625	0.524	0.487	0.347	0.464	0.373	0.439
Qwen3-VL-235B-A22B	**0.708**	**0.693**	0.825	0.645	0.457	**0.695**	**0.944**	**0.484**	**0.672**	**0.577**	**0.547**	**0.421**	**0.519**	**0.433**	**0.508**
Open-source Thinking VLMs															
Keye-VL-1.5-8B	0.632	0.675	0.702	0.526	0.527	0.641	0.959	0.553	0.72	0.622	0.614	0.432	0.532	0.441	0.53
InternVL3.5-8B	0.576	0.74	0.553	0.453	0.483	0.688	0.983	0.585	0.744	0.655	0.649	0.423	0.534	0.446	0.533
Qwen3-VL-8B	0.607	0.753	0.586	0.489	0.534	0.637	0.955	0.606	0.737	0.648	0.659	0.482	0.572	0.492	0.575
Qwen3-VL-32B	**0.677**	0.741	**0.716**	**0.587**	**0.547**	**0.622**	0.919	**0.646**	0.755	0.674	0.685	**0.558**	**0.625**	**0.549**	**0.639**
InternVL3.5-38B	0.613	0.688	0.677	0.518	0.515	**0.622**	**0.907**	0.604	**0.764**	**0.679**	0.667	0.444	0.551	0.462	0.55
Qwen3-VL-235B-A22B	0.646	**0.755**	0.637	0.54	0.516	0.639	0.918	0.637	0.761	0.678	**0.687**	0.524	0.61	0.526	0.616
Closed-source Models (Thinking Minimized)															
GPT-5	0.647	0.703	0.678	0.571	0.427	0.737	0.981	0.483	0.673	0.575	0.542	0.449	0.563	0.474	0.526
Gemini-2.5-Pro	**0.697**	**0.718**	**0.762**	**0.626**	**0.549**	**0.578**	**0.851**	**0.6**	**0.76**	**0.671**	**0.653**	**0.513**	**0.593**	**0.51**	**0.581**
Closed-source Models (Full Thinking)															
GPT-5	0.653	**0.734**	0.656	0.566	**0.566**	**0.587**	**0.879**	**0.661**	**0.799**	**0.72**	**0.698**	**0.529**	**0.609**	**0.529**	**0.603**
Gemini-2.5-Pro	**0.69**	0.718	**0.751**	**0.58**	0.556	0.59	0.885	0.651	0.792	0.711	0.696	0.511	0.591	0.512	0.59
Caption Baseline															
Gemini-Caption-S	0.687	**0.737**	0.72	0.599	0.572	0.573	0.874	0.557	0.734	0.64	0.61	0.468	0.584	0.495	0.549
Gemini-Caption-L	**0.703**	0.735	0.757	**0.65**	0.58	**0.544**	**0.829**	0.585	0.743	0.654	0.634	0.498	0.592	0.505	0.574
Gemini-Caption-S-T	0.68	0.712	0.733	0.574	0.572	0.599	0.916	0.584	0.758	0.669	0.636	0.485	0.58	0.495	0.557
Gemini-Caption-L-T	0.695	0.722	**0.759**	0.594	**0.586**	0.568	0.875	**0.642**	**0.786**	**0.703**	**0.685**	**0.524**	**0.611**	**0.531**	**0.597**

Table 1. Task results when *visual context* is provided, meaning all images in the album were given to the VLM. Bold indicates the best performance in the given partition, and underline indicates the best performance overall. In the model names for the "Caption Baseline" partition, "S" means a short caption was provided, "L" means a long caption was provided, and "T" means the thinking mode was used.

Model	Intent Selection F1↑	Prec.↑	Recall↑	mAP↑	Intent Rating Acc.↑	MAE↓	RMSE↓	Group Labeling ARI↑	F1↑	Jaccard↑	NMI↑
Open-source Instruct VLMs											
Keye-VL-1.5-8B	0.624	0.506	**0.964**	0.558	0.295	0.893	1.075	0.517	0.666	0.58	0.585
InternVL3.5-8B	0.653	0.552	0.939	0.587	0.331	0.848	1.055	0.531	0.675	0.59	0.588
Qwen3-VL-8B	0.661	0.57	0.92	0.578	0.298	0.872	1.058	0.565	0.707	0.623	0.622
Qwen3-VL-32B	0.688	0.621	0.889	0.608	0.397	**0.76**	**0.989**	0.597	0.735	0.652	0.645
InternVL3.5-38B	0.683	0.604	0.913	0.606	0.344	0.867	1.098	0.59	0.74	0.654	0.64
Qwen3-VL-235B-A22B	**0.691**	**0.629**	0.888	**0.611**	**0.403**	0.77	1.009	**0.615**	**0.758**	**0.676**	**0.661**
Closed-source Models (Thinking Minimized)											
GPT-5	**0.68**	**0.633**	0.849	**0.597**	0.503	**0.648**	**0.915**	0.613	0.748	0.666	0.656
Gemini-2.5-Pro	0.678	0.611	**0.893**	0.592	**0.517**	0.669	0.962	**0.614**	**0.755**	**0.672**	**0.659**
Caption Baseline											
Gemini-Caption-S	0.673	0.612	**0.863**	0.58	0.485	0.714	1.01	0.55	0.725	0.633	0.596
Gemini-Caption-L	**0.676**	**0.626**	0.851	**0.592**	**0.521**	**0.664**	**0.965**	**0.613**	**0.754**	**0.67**	**0.657**

Table 2. Task results when *language context* is provided, meaning a caption was used to provide context for the album. Bold indicates the best performance in the given partition, and underline indicates the best performance overall. In the model names for the "Caption Baseline" partition, "S" means a short caption was provided, "L" means a long caption was provided.

Discussion

Open-source vs. Closed-source:

- Qwen3-VL generally better among open-source models
- Gemini-2.5-Pro & GPT-5 each takes the lead on half the tasks

How well do VLMs understand the joint context of images?

- All models struggle with understanding joint context of albums
- "Thinking" greatly improves performance on grouping tasks

Do VLMs make good use of visual context?

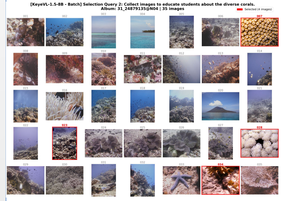
- Gemini-Caption-Baselines comparable to Gemini-2.5-Pro
- Group Labeling: language context >> visual context
- Visual information is **underutilized**

Failure modes:

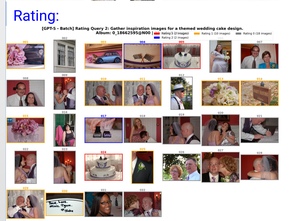
- Missing images and overlapping groups
  - seen frequently on smaller and open-source models
- Model gives empty answers
  - Qwen-3-VL and Gemini-2.5-Pro
- Over-simplified categories and conceptually overlapped categories
  - Keye-VL-1.5 generates two groups: "Wedding Scenes" and "Misc." for a wedding

Examples

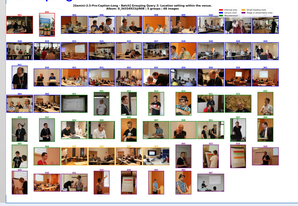
Selection:



Rating:



Grouping:



Contact

Shawn Huang
Brigham Young University
Email: huang717@byu.edu

Project Link

https://byu-vision.github.io/albumbench/

References

1. https://photutorial.com/photos-statistics/
2. Y. Wang, Z. Lin, X. Shen, R. Mech, G. Miller and G. W. Cottrell, "Event-Specific Image Importance," 2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), Las Vegas, NV, USA, 2016, pp. 4810-4819, doi: 10.1109/CVPR.2016.520.
3. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., ... Chen, W. (2024). MMMU: A massive multi-discipline multimodal understanding and reasoning benchmark for expert AGI. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 17134–17145.
4. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al. (2025). Video-MME: The first-ever comprehensive evaluation benchmark of multi-modal LLMs in video analysis. CVPR.
5. Song, D., Chen, S., Chen, G. H., Yu, F., Wan, X., & Wang, B. (2024). MileBench: Benchmarking MLLMs in long context. arXiv preprint arXiv:2404.18532.
6. Wang, F., Fu, X., Huang, J. Y., Li, Z., Liu, Q., Liu, X., ... Chen, M. (2025). MuirBench: A comprehensive benchmark for robust multi-image understanding. The Thirteenth International Conference on Learning Representations. https://openreview.net/forum?id=TrVYEzSQH
7. Lin, TY. et al. (2014). Microsoft COCO: Common Objects in Context. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds) Computer Vision – ECCV 2014. ECCV 2014. Lecture Notes in Computer Science, vol 8693. Springer, Cham. https://doi.org/10.1007/978-3-319-10602-1_48
8. Li, Y., Wu, S., Gao, Z., Zhang, Z., Jia, Q., Duan, H., Min, X., & Zhai, G. (2026). VideoAesBench: Benchmarking the video aesthetics perception capabilities of large multimodal models. arXiv preprint arXiv:2601.21915. https://arxiv.org/abs/2601.21915
9. Hudson, D. A., & Manning, C. D. (2019). GQA: A new dataset for real-world visual reasoning and compositional question answering. Conference on Computer Vision and Pattern Recognition (CVPR).
10. N. Murray, L. Marchesotti and F. Perronnin, "AVA: A large-scale database for aesthetic visual analysis," 2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, 2012, pp. 2408-2415, doi: 10.1109/CVPR.2012.6247954.

© POSTER TEMPLATE BY GENIGRAPHICS® 1.800.790.4001 WWW.GENIGRAPHICS.COM