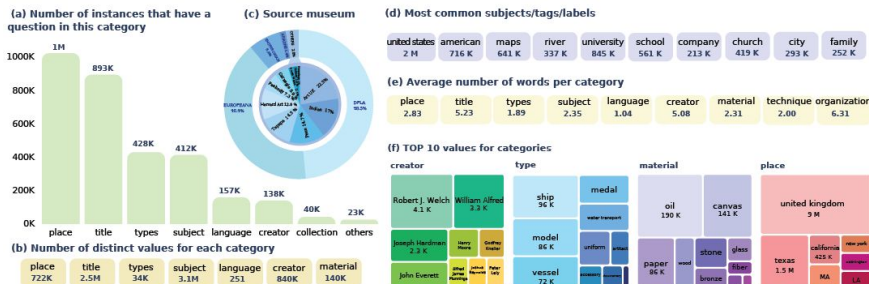


Dataset



- **Large-scale and diverse:** Contains **65 million images** and **200 million question-answer pairs** spanning art, history, and natural sciences, covering exhibits from museums worldwide.
- **Expert-curated & multilingual:** Data is cleaned and labeled by museum experts, featuring questions in **38 languages** (English + 37 more)
- **Structured for reasoning:** information of each exhibit is in the form of **attribute-value pairs** (e.g., title, material, place, creator) designed for VQA tasks to test **domain-specific understanding**.



Code & Contact

Code is available open-source
[GitHub - institut-institut/Museum-65](https://github.com/institut-institut/Museum-65)
Contact: Sanjana Garai
sanjana.garai.mat22@itbhu.ac.in



Conclusion

Fine-tuning on MUSEUM-65 empowers vision-language models with superior accuracy and reasoning, proving its utility for cultural heritage AI.

Tasks

Task 1: General VQA

Q: What is the medium of the object?
A: oil on canvas

LLaVA20mm: oil on canvas
BLIP20mm: oil on canvas

Q: What is the title of this object?
A: silver denarius (empire)

LLaVA20mm: silver denarius (empire)
BLIP20mm: silver penny

Task 3: MultiAngle

Q: Who is the artist of the object?
A: David Nash

LLaVA20mm: David Nash
BLIP20mm: unknown artist

Q: What are the subjects of the object?
A: Commemorative, Military Non-figurative, Second

LLaVA20mm: Commemorative, Non-figurative
BLIP20mm: war memorial

Task 4: Visually Unanswerable Questions

Q: What is the form of government in the country of origin of this object?
A: Parliamentary Republic

LLaVA20mm: Republic
BLIP20mm: Italy

Q: Which part did the country of origin of this object support during WW2?
A: Neutral, occupied by Nazi Germany

LLaVA20mm: Germany
BLIP20mm: Denmark

Task 5: MultiLanguage

Q: Quel est le titre de l'objet?
A: Chemins de fer de l'Ouest, Gare St Lazare, Stréat à 4 heures de Paris, Tennis club

LLaVA20mm: Etréat à 4 nuits de Paris
BLIP20mm: United States

Q: Geben Sie eine kurze Beschreibung des Objekts?
A: Partitur, München, Bayerische Staatsbibliothek

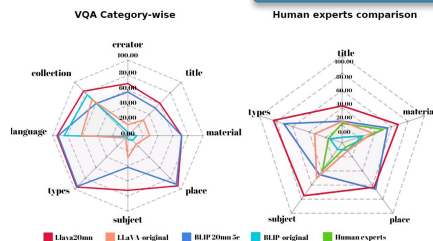
LLaVA20mm: Partitur
BLIP20mm: United States

Model	General VQA	Multiangle		Multilang		Visually Unanswerable
		Original	Alternate	French	German	
LLaVA 20mm 1 ep	80	58.09	56.14	10.37	9.72	29.7
LLaVA nofinetune	23	24.35	23.56	41.81	18.41	27.18
BLIP 20mm 5 ep	72	52.78	51.75	4.02	0.73	3.3
BLIP nofinetune	9	13.82	12.86	2.01	0.8	5.44

Table 1: Performance of LLaVA and BLIP models across multiple tasks.

The five tasks evaluate models on **general understanding**, **attribute reasoning**, **viewpoint robustness**, **contextual inference**, and **multilingual comprehension**. Results show that **fine-tuning on MUSEUM-65** greatly boosts performance across all tasks, with LLaVA excelling in complex reasoning and multilingual settings.

Other experiments



Model	Europe	N. America	S. America	Asia	Africa	Oceania
LLaVA 20mm-1ep	85.2	79.6	86.6	67.4	86.7	99.2
LLaVA	8.6	43.57	20.3	23.4	20.79	52.4
BLIP-20mm-5ep	79.1	73.1	76.4	65.5	76.4	49.7
BLIP	4.3	15.2	19.7	9.3	19.7	6.6

Table 2: Continent-wise Partial Precision.

Continent-wise partial precision shows that **fine-tuning on our domain-specific dataset** enables consistent gains across all regions, demonstrating robustness despite data imbalance.

Model	BLIP nofinetune	BLIP 20mm 5e	LLaVA nofinetune	LLaVA 20mm 1e
METEOR	3.24	37.45	2.96	58.85
WMD Acc.	35.54	74.02	54.5	87.02

Table 3: Semantic Evaluation results.

LLaVA-20mm, consistently outperform original models and human experts across most categories, demonstrating **superior domain-specific reasoning**.

Fine-tuning greatly enhances semantic understanding.