HierArtEx: Hierarchical Representations and Art Experts Supporting the Retrieval of Museums in the Metaverse

Alex Falcon[1,*][0000−0002−6325−9066], Ali Abdari[1,2][0000−0002−4482−0479], and Giuseppe Serra[1][0000−0002−4269−4501]

[1] University of Udine, Italy
[2] University of Naples Federico II, Italy

Abstract. The improvements in Virtual Reality technologies are bringing more attention to the Metaverse and the nearly unlimited experiences available there. Among these, digital museums have seen an increase in the number of yearly visitors, especially after the COVID-19 pandemics. However, no tools are available to support the user in the searching process. To this end, we start investigating the Text-to-Museum retrieval task, involving museums composed of many rooms enriched by multimedia elements affecting their relevance to the user query. To model this complex type of data, we design HierArtEx, which leverages hierarchical representations to model the whole museum, while combining generic and art-specific knowledge for capturing the visual contents of each single room. We validate its effectiveness on Museums3k, a large dataset that we collect, containing 3000 realistic museums each annotated by a description of its contents. Moreover, qualitative analyses confirm favorable results and their alignment with real user queries, while also highlighting the shortcomings of standard evaluation protocols in retrieval, as they fail to capture all relevant museums.

Keywords: Multimedia Retrieval · Cross-modal Retrieval · Metaverse · Complex 3D Scenarios · Digital Museums.

1 Introduction

The concept of "digital museum" has been studied since the advent of the Internet, due to the new possibilities envisioned for making the heritage widely accessible and bringing the contents of traditional museums to broader audiences [13]. Nowadays, many traditional museums offer digital immersive experiences, e.g. the Smithsonian American Art Museum, the Louvre, and the National Museum of Finland, among many others. These experiences are often implemented using Virtual Reality (VR) technology, in which the users feel truly immersed and report improved engagement, realism, and satisfaction [5, 21]. Moreover, the use of VR leads to increased educational value and effectiveness of the experiences [18]. Therefore, the trend is for more traditional museums to implement

* Corresponding author: Alex Falcon (falcon.alex@spes.uniud.it)

their own digital experiences in the Metaverse. However, this process will lead to enormous collections of Metaverse museums resulting in the difficulty of filtering them to find the few which fit the current interests of the user. For instance, the Ministry of Culture and Tourism of China reported that more than 2000 exhibitions were held online and that they attracted over 5 billion international visitors [10]. Therefore, intelligent software tools to support users during the search process are becoming an urgent need.

A key step to implementing an intelligent tool for text-to-museum retrieval consists in modeling the museum itself. In fact, Metaverse museums can be seen as 3D scenarios filled with multimedia elements, such as paintings on the walls, artifacts stored in frameless display cases, and sculptures often located at the center of the room so that visitors are free to walk around them and view them from different angles. Current literature on modeling such complex scenarios is lacking, as the focus is usually put on text-based retrieval of single 3D objects [11, 16, 25], with few works dedicated to retrieving indoor scenarios containing multiple 3D objects [1, 3, 24]. As multimedia elements are an important part of the Metaverse scenarios, since they influence the relevance to the user query, then even fewer works can be found, mostly dedicated to implementing Text-to-Metaverse Retrieval by realizing proof-of-concept datasets with paintings stitched to existing indoor scenarios [2, 4].

Therefore, we highlight two intertwined shortcomings in the literature: the lack of suitable datasets for investigating text-guided retrieval of Metaverse museums, and the lack of approaches to model complex 3D scenarios containing both multiple rooms and many multimedia elements which affect the relevance to the user query. In this paper, we address the first shortcoming by collecting a large dataset of 3000 realistic museums containing multiple rooms decorated with paintings on the walls and annotating them with a description of their contents, room-by-room. Then, we present a methodology, called HierArtEx, a simple yet effective solution based on CLIP [19] to perform text-to-museum retrieval. With HierArtEx, the aim is to capture the hierarchical structure of the museum by explicitly modeling it, while the content of the paintings is captured by leveraging both generic and specific knowledge obtained through pretrained experts. The experimental results show the effectiveness of HierArtEx in performing Text-to-Museum retrieval, and the ablation study confirms the usefulness of each component introduced in this work. Moreover, qualitative analysis is performed to understand the retrieval results and how they may align with user intents.

The main contributions of this paper are summarized as follows:

- We highlight the lack of available methodologies for retrieving digital museums using textual queries. This is becoming more crucial as Metaverse technologies improve over time and more users begin to use them. To address this shortcoming, we propose HierArtEx, a methodology for capturing the hierarchical structure of the museums while also leveraging pretrained experts, both generic and art-specific, to understand their contents.
- We collect a large-scale dataset, called Museums3k, comprising 3000 museums, represented as 3D scenarios containing many rooms and multimedia

elements in each of them. We evaluate HierArtEx on Museums3k, confirming its effectiveness and the usefulness of each component through ablation studies.
– We qualitatively investigate the retrieval performance of HierArtEx, highlighting a shortcoming in standard evaluation protocols, as recall metrics fail at capturing the complexity of the museums and therefore at measuring the retrieval of near-correct museums. This also highlights possible future research directions specific to the Text-to-Museum retrieval problem.

The rest of the paper is organized as follows. In Section 2, a review of the literature is performed to identify the related works. Details on how we collected the dataset are contained in Section 3, whereas our methodology is explained in Section 4. The experimental results are presented in Section 5, whereas the limitations of this work are described in Section 6. Finally, Section 7 concludes the paper.

2 Related Work

2.1 Digital Museums

The development of digital museums has been a longstanding field due to its potential of bringing cultural heritage to a broader audience, while drastically reducing the costs related to cultural tourism. This field raised even more interest in recent years, especially during the COVID-19 pandemic, as physical accesses to museums were restricted, resulting in a considerable decrease in yearly visits [12]. With recent advances in technology, it has become easier to build highly detailed 3D models of tangible heritage coming from varied topics, leading to the creation of digital museums on archaeological heritage [6, 17], but also on modern heritage, e.g. naval engineering [7]. To support the creation of virtual museums, often done manually after the digitization of the tangible heritage, many tools with a focus on VR technology were developed [14, 26]. In fact, while Augmented Reality allows for better user experience during physical visits, VR represents the primary method for implementing fully virtual museums, with the users feeling satisfied with the experience [5], and reporting high levels of presence, engagement, and immersion [21], and also enhanced effect on learning [18]. In summary, these techniques and advancements led to an increasing trend in creating immersive museum experiences in the Metaverse.

In this work, to establish a foundation for research on large-scale analysis of Metaverse museums, we collect a dataset comprising 3000 virtual museums, each containing multiple rooms furnished with several paintings located on the walls. An example is shown in Fig. 1.

2.2 Text-guided Retrieval of 3D Scenario

Recently, cross-modal approaches for retrieving 3D objects have attracted a lot of interest from the research community, especially after CLIP [19]. Le et al. used

In this museum there are six rooms. In the first room, there are four paintings. There is one painting from the French school with a landscape style, one painting from the Flemish school with a religious style, [...]. One painting is titled Water Lilies [...]. From 1899 to 1926 Monet worked on the theme of the water lilies with unremitting determination, using different formats and viewpoints, adopting an increasingly flexible treatment, [...] One painting is titled Rue Montorgueil in Paris, Celebration of 30 June 1878 [...] a sea of flags was fluttering in the wind, and countless people filled the streets. [...] gives the illusion of an ethereal carpet of colour. [...]

Fig. 1. An example from the collected dataset. On the left, the 3D museum is shown from above with isometric perspective. On the right, three rooms are highlighted to show the paintings on the walls. The description of the museum is put below.

CLIP as the encoder both for the textual queries and the images taken at the 3D objects from multiple angles [15]. Xue et al. integrated the 2D knowledge of CLIP with 3D knowledge from point clouds [23], whereas Hegde et al. incorporated prompt tuning and additional training objectives [11]. Liu et al. further extends the training data scale achieving better results on standard and new benchmarks for 3D object retrieval [16]. However, all these studies investigated the retrieval of *single* 3D objects. Recently, the retrieval of richer 3D scenarios, which include *multiple* 3D objects which affect the relevance to the textual query, has seen the first works that started to investigate this problem on specific settings, such as furnished indoor scenes [1, 3, 24] and multimedia-rich 3D scenes [2, 4].

In this work, we focus on the latter problem, that is the retrieval of rich 3D scenarios. In particular, we focus on the retrieval of digital museums, i.e. 3D scenarios comprising multiple rooms and many multimedia elements. Moreover, we collect a novel, large scale dataset of 3000 realistic museums.

3 Museums3k: A Digital Museums Dataset

To address the lack of datasets of digital museums, we designed a procedure to procedurally create them using Unity. Specifically, the procedure is done in two steps. First, the number of rooms to be created for the museum is selected. In

our dataset, this number ranges from six to nine. Then, the procedure creates an empty room with one door and starts inserting one room after the other with doors connecting subsequent rooms (see left side of Figure 1 for an example). As a second step, paintings are put in every empty room, starting from the second one, while the first room is kept as an empty lobby. In particular, two paintings are put on every wall that does not contain a door, so that every room has around four paintings. As a source for the paintings and for the descriptions forming our textual annotations, we use SemArt [9], which contains more than 21000 samples annotated with a textual description. As can be seen in the example, these descriptions provide a comment on the historical background ("from 1899 to 1926 Monet worked on the theme of the water lilies with unremitting determination," etc) or on the visual contents ("a sea of flags was fluttering in the wind, and countless people filled the streets. [...] gives the illusion of an ethereal carpet of colour."). One of the museums created in this way is shown in Figure 1. Note that the first room (lobby room) is ignored, as it does not contain any multimedia element, and so only the decorated rooms are considered, hence why the description of the example says "six" rooms instead of "seven". In total, we collected 3000 museums, containing about 28 paintings per museum resulting in a total of about 12M tokens (on average, 3983 tokens and 170 sentences per museum).

4 Proposed Methodology

Figure 2 presents an overview of the proposed method, called HierArtEx, for solving the Text-to-Metaverse retrieval problem by leveraging a hierarchical representation for the museums supported by an art "expert". It is made of three main components: a module for learning image-level representations supported by an art expert (Sec. 4.1); a hierarchical representation module for learning museum-level representations (Sec. 4.2); and, finally, a module dedicated to processing the museum descriptions (Sec. 4.3). The pipeline can be described as follows. First, from each room of every museum, a set of twelve images is captured by placing a camera in the center of the room and rotating it by 30 degrees at a time. These are then processed using a combination of two methods, in order to obtain a good mixture of generic and art-specific knowledge. For the former, we use CLIP [19], whereas for the latter, we use ArtExp, a standard vision backbone that we pretrain using multi-task learning on a large dataset of painting-genre-style triplets. Painting-level representations are then obtained by transforming the concatenated generic-specific knowledge through $f_{painting}$. Then, painting-level representations are put together through f_{room} to obtain a representation of the single room and all the paintings contained in it. Finally, to obtain a museum-level representation, the room representations are transformed through f_{museum}. To ease the museum-text alignment, fundamental for implementing a cross-modal retrieval system, the museums' descriptions are also processed. To do so, they are first split into sentences, as the descriptions tend to be very long due to the abundance of paintings available in each

museum. Then, sentence-level representations are extracted using the textual encoder of CLIP, and finally, description-level representations are obtained by means of $g_{description}$, implemented using a bidirectional GRU. These functions, $f_{painting}$, f_{room}, f_{museum}, and $g_{description}$ are jointly learned at training time in a contrastive manner (Sec. 4.4).

Further details are presented in the following sections. Section 4.1 introduces the procedure of obtaining generic and specific knowledge from the images. Section 4.2 provides details of the hierarchical representations that we learn for the Metaverse museums. The processing of the textual descriptions is discussed in Section 4.3. Finally, the training procedure of the entire retrieval model is explained in Section 4.4.

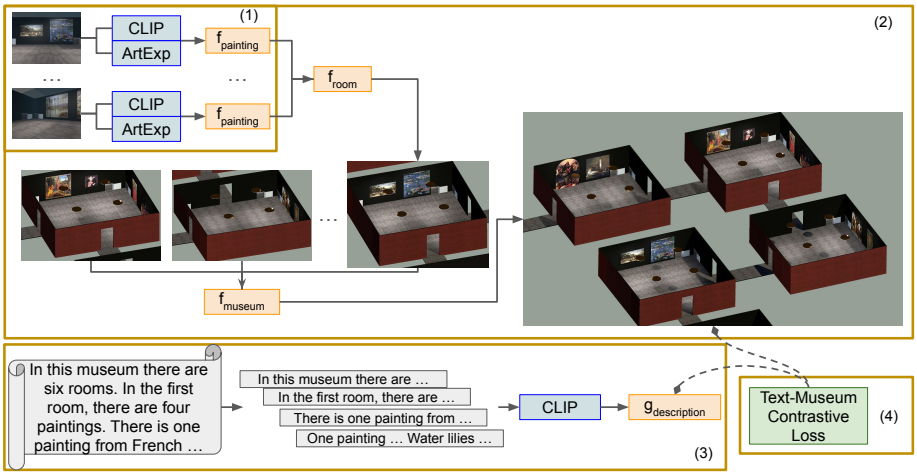

Fig. 2. Overview of the presented method, HierArtEx. It is made of three main components. (1) Image-level representations are obtained using a mix of generic (CLIP) and specific (ArtExp) knowledge from pretrained models, and then combined by $f_{painting}$. (2) A museum-level representation is obtained in two steps, by first combining the image-level representations using f_{room} and then by combining the room-level representations using f_{museum}. (3) The textual description of the museum is first split into sentences, and then processed using CLIP and $g_{description}$. (4) Contrastive learning is used to jointly adjust the weights of the trainable functions from both modalities.

4.1 Obtaining generic and specific knowledge from the images

From each room in the museum, multiple images are extracted. Then, each image is processed using two different pretrained models. First, CLIP is used as a generalist vision-language model, the features of which are related to general aspects of the visual contents, since CLIP was pretrained on large amounts of generic image-caption pairs scraped from the web. This results in a vector

$x^i_{gen} \in R^{1\times D_{gen}}$. Second, ArtExp acts as the art expert, since it is a model which we pretrained using multi-task learning to detect multiple artistic information (e.g. genre and style) from each painting (details in Section 5.1). With ArtExp, a vector $x^i_{art} \in R^{1\times D_{art}}$ is obtained. Then, we concatenate these two vectors to obtain $x^i_{ga} \in R^{1\times D_{gen}+D_{art}}$. This vector is then processed through $f_{painting}$, a trainable function that we implement with a linear transformation and a non-linear ReLU activation function. The final representation for each image is a vector $x^i_{pic} \in R^{1\times D_{ptn}}$.

4.2 Hierarchical representation for Metaverse museums

To obtain the features for room j, the representations of the images contained in such room are processed and aggregated. To do so, we first take the mean of the image vectors $x^i_{pic}, i = 1, \ldots, N^j_{pic}$, and then process them through the trainable function f_{room}, obtaining $r_j \in R^{1\times D_{room}}$. Here, N^j_{pic} is the number of pictures taken from room j. As in the previous case, f_{room} is implemented with a linear transformation and a ReLU. Then, the representations of all the rooms in the museum are computed and processed through similar steps to obtain the museum-level representation. That is, we take the mean of $r_j, j = 1, \ldots, N^k_{room}$, where N^k_{room} is the number of rooms for museum k, and then process it through f_{museum}, implemented as a linear transformation, resulting in the final vector $m_k \in R^{1\times D_{mus}}$.

4.3 Processing the museum descriptions

As mentioned before, the descriptions of the museums are fairly long, as they need to describe every painting found in the museum. This because each of them may contribute to the relevance to the user query, e.g. if it is about museums with religious paintings from Italian Renaissance period then having one or more of them will make the museum more relevant. To process them and understand their contents, we split the descriptions into sentences and then process each sentence through the textual encoder of CLIP, obtaining $d^i_k \in R^{1\times D_{sent}}, i = 1, \ldots, N^k_{sent}$ where N^k_{sent} is the number of sentences for the description of museum k. The list of sentences is then processed through the function $g_{description}$, implemented as a bidirectional GRU followed by the mean of the last hidden state from each direction, to obtain $t_k \in R^{1\times D_{txt}}$. A similar procedure was also followed in previous works where descriptions were quite long [1, 3].

4.4 Training procedure of HierArtEx

The training procedure aims at aligning the information obtained from the museum to that obtained from the textual descriptions. To achieve this goal, we compute the triplet loss [20] (Eq. 1) on all negatives. This means that given a list of museum-level representations $m_k, k = 1, \ldots, B$ and the corresponding descriptions $t_k, k = 1, \ldots, B$, the loss $\mathcal{L}$ is computed as following:

$$L(a,p,n) = max(0, \Delta + sim(a,n) - sim(a,p)) \tag{1}$$

$$\mathcal{L} = \frac{1}{2B^2}\left(\sum_{i=1}^{B}\sum_{j=1,j\neq i}^{B} L(m_i,t_i,t_j) + \sum_{i=1}^{B}\sum_{j=1,j\neq i}^{B} L(t_i,m_i,m_j)\right) \tag{2}$$

where Δ is a fixed margin hyperparameter, sim is the cosine similarity, and B is the batch size. By optimizing with Eq. 2, the model is taught to output similar representations for m_i and t_i, while keeping a distance of Δ between their cosine similarity and that of m_i (or t_i) and a negative t_j (or m_j), i.e. a description (or museum) not paired to m_i (or t_i) in the dataset.

5 Experimental results

5.1 Implementation details

As mentioned in Sec. 4, in each museum, we put a camera in the middle of each decorated room and extract 12 images separated by 30 degrees difference. The 3000 obtained museums are split into 2100 (70%), 450 (15%), and 450 (15%) for training, validation, and testing, respectively. All the results are obtained by performing the training three times, and then the average performance on the test set is reported after selecting the best model on the validation set.

We train HierArtEx for 50 epochs using a batch size B of 64, and the Adam optimizer with a learning rate of 1e-3 and a gamma of 0.75 after 27 epochs. The experiments are performed on a machine running a single A5000 GPU and Intel Xeon W-2123 CPU (3.60GHz). PyTorch 1.13.1 is used as the deep learning framework. The sizes D_{gen} and D_{sent} are 512 as this is the CLIP output size. D_{art} depends on the backbone: it is 512 for ResNet18 and ResNet34, 2048 for ResNet50 and ResNet101, and 768 for ViT-B-16 and ViT-B-32. D_{ptn}, D_{room}, D_{mus}, and D_{txt} are all set to 256. In our setting, the number of pictures taken in each room N^j_{pic} is set to 12. Δ is set to 0.25.

The art expert is trained on a machine running one A100 GPU coupled with AMD EPYC 7643 48-Core CPU. The dataset used for training is a subset of WikiArt [22], obtained by keeping only the paintings for which both genre and style labels are available, which leads to about 65000 paintings. Since the classes are not balanced (e.g. ranging from 96 to more than 11000 paintings per style), we select 80% of the paintings grouped by the style class as a training set (51997) and leave the rest (12998) for validation. The training lasts 25 epochs, with two heads attached on top of the backbone learning to predict style and genre in a multi-task learning setting. Random crop and horizontal flip are applied during training, whereas central crop and resize are performed during validation. The SGD optimizer was used, with a learning rate set to 2e-5, momentum of 0.9, and gamma of 0.1 every 7 epochs. The batch size is 32. Here, we used PyTorch 2.1.2 and torchvision 0.16.2.

Code and data are available at https://github.com/aranciokov/HierArtEx-MMM2025.

Table 1. Results of the ablation study of the proposed method, HierArtEx. Discussion in Section 5.2.

Method	Text-to-Museum				Museum-to-Text			
	R1	R5	R10	MedR	R1	R5	R10	MedR
HierArtEx	47.7	81.7	90.8	2.0	44.3	80.4	89.6	2.0
w/o ArtExp	36.6	70.9	82.1	2.3	36.1	68.3	81.0	2.3
w/o ArtExp and Hier (=Baseline)	14.2	40.4	56.7	8.7	14.4	37.9	54.0	9.5

5.2 Ablation study and performance evaluation

Here we evaluate the performance of the proposed method, HierArtEx, on the Text-to-Museum retrieval task. Alongside these results, Table 1 also contains an ablation study on the novel components added in HierArtEx, that is the hierarchical representation learning and the usage of art experts to enrich the image-level representations. The proposed method achieves 47.7% R@1 and 90.8% R@10, indicating its great capability in identifying a museum given its description. When the art expert is removed, there is a drop in performance, leading to 36.6% R@1 and 82.1% R@10, a drop of -11.1% and -8.7%, respectively. This indicates that supporting the image-level representations with specific knowledge from art experts is useful in identifying the museum more precisely. If we further remove the hierarchical representations, the baseline is obtained. In this case, the performance drop is considerable, as almost -30% is observed in both R@1 and R@10, leading to 14.2% and 56.7%, respectively. These results show the effectiveness of both components and their synergy with the baseline to obtain the proposed method, HierArtEx.

5.3 On the design of ArtExp

As mentioned in Section 5.1, we used standard vision backbones for the art expert and trained it using a multi-task learning strategy. Here, we performed an experiment to identify how different backbones performed both on the custom WikiArt subset and on the Museums dataset that we collected. The results are reported in Table 2.

We chose ResNet in its 18, 34, 50, and 101 versions, and Vision Transformers [8] in the ViT-B-16 and ViT-B-32 implementations from the torchvision library. The experimental results show that the best results on Museums3k are achieved by implementing ArtExp as either ResNet50 or ResNet101, obtaining up to 47.7% R@1. Interestingly, these results do not seem to correlate with stronger performance on the WikiArt dataset, as ResNet34 and ViT-B-16 achieve the best genre and style accuracy, measuring up to 71.8%/47.0% and 72.5%/51.5%, respectively, while their performance on Museums3k is lower (about -20% R@1 compared to ResNet50). While the initial pretrained weights do not seem to matter, as ResNet50, ResNet101, and ViT-B-16 all achieve about 81% accuracy on ImageNet, the number of parameters and the size of the learned embeddings may represent two hints at better transferring of art-specific knowledge.

Table 2. Experiment with different backbones for the implementation of ArtExp. Performance on the Museums3k dataset is reported for the text-to-museum direction, and genre and style accuracy is reported for WikiArt. ImageNet (IN) accuracy, number of parameters (in millions), and size of the learned embeddings are also reported for the backbones. Discussion in Section 5.3.

Art Backbone	Text-to-Museum				Genre	Style	IN	Num	Size
	R1	R5	R10	MedR	Acc	Acc	Acc	Params	Embeds
ResNet18	22.1	52.6	66.8	5.3	65.8	44.1	69.7	11.7	512
ResNet34	27.8	57.6	70.5	3.7	71.8	47.0	73.3	21.8	512
ResNet50	**47.7**	**81.7**	**90.8**	**2.0**	65.9	34.5	80.8	25.6	2048
ResNet101	40.4	73.7	85.4	2.3	64.9	34.5	81.9	44.5	2048
ViT-B-16	26.7	58.3	72.0	3.8	**72.5**	**51.5**	81.1	86.6	768
ViT-B-32	17.2	45.9	61.3	6.3	66.6	41.7	75.9	88.2	768

5.4 Qualitative analysis of the retrieval results

Verifying that a retrieved museum is indeed relevant to the query is not a straightforward task. In the previous sections, we followed standard procedures used in other cross-modal retrieval settings and measured how many times the museum was retrieved in the top 1/5/10 of the ranked list given its description as a query. However, there are two key issues with these protocols. First, there may be similarities among different museums which makes them somehow relevant to other museums, and these similarities are not taken into account when computing the recall metrics. Second, descriptions can be very different compared to user queries: a user query is likely much simpler, shorter, and more vague, e.g. "museum focusing on Impressionist painters" or "museum on the art of war, with paintings ranging from the training to the final battle". In this analysis, we try to capture this behavior by working on the idea of "common concepts in the museum", which we consider as those words in the related descriptions which are repeated many times and are meaningful and descriptive of the visual contents, i.e. excluding non-stop words, numbers, and other narrative words ("however", "although", but also "described", "titled", etc). For instance, we keep generic words related to colors, environment (e.g. trees, walls, and bridge), and foreground elements (e.g. angels, cathedral, and dragon), and more specific words related to style (e.g. religious, landscape, and portrait), important characters (e.g. Jesus, Cecilia, and Jerome), and events (e.g. Crucifixion and Annunciation). In total, we kept from about 150 to 700 unique concepts when looking for 5 to 15 most common concepts. An example of these concepts, from example queries and retrieved museums from the test set, is shown in Fig. 3. Note that the concepts are only used for visualization purposes, and do not affect the retrieval in any way.

Then, we use these concepts to compare HierArtEx and the baseline over all the test samples. For each method and for each test query, we retrieve the top 1/5/10 museums and obtain a score determined by summing the number of concepts found in those museums which are also part of the "groundtruth concepts", i.e. those of the query. Then, we keep track of how many times HierArtEx ob-

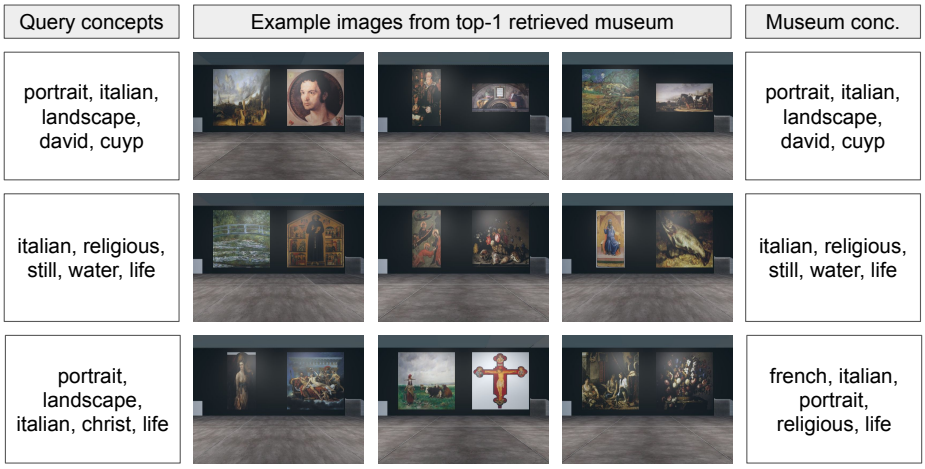

Fig. 3. Examples of the important "concepts" in the query and those found in retrieved museums (relevant images are shown in the middle). Discussion in Sec. 5.4.

Table 3. Each entry indicates the number of times HierArtEx retrieved more relevant museums than the baseline, based on a score computed through the analysis of important "concepts" of the query. Discussion in Sec. 5.4.

Retrieved Museums	Concepts		
	5	10	15
Top-1	228/116/106	234/102/114	233/91/126
Top-5	230/74/146	262/45/143	271/28/151
Top-10	223/37/190	222/38/190	253/26/171

tains a higher/equal/lesser score than the baseline. Results are reported in Table 3, where each entry indicates the three numbers. The results further confirm the effectiveness of HierArtEx, which consistently retrieves more relevant museums according to the concepts, both when increasing the number of concepts to be considered (left to right), and when considering more museums in the evaluation (top to bottom). Moreover, another observation can be made which highlights the complexity of evaluating Text-to-Museum retrieval methods and its relation with the issues mentioned at the beginning of this section. In the third row of Fig. 3, the top-1 retrieved museum by HierArtEx is not the correct one. Yet, both the concepts of that museum and the paintings contained in it, show that this museum could likely be a good result for the user: in fact, there are both portraits and landscapes, still-life paintings, and religious paintings, with Christ shown on the cross in the second image.

6 Discussion and future work

The final goal of this research lies in supporting the user in searching interesting museums to be explored in the Metaverse. In Sections 5.2 and 5.3, we show our proposed method performs better than the CLIP-based baseline we considered. However, two key considerations need to be taken. First, in Museums3k, the museums and even the rooms are not necessarily thematic and the visiting experience is not manually designed by an expert, limiting their similarity to real museums. Future work should assess the proposed method in more realistic scenarios. Second, in Section 5.4, we introduce a way to qualitatively analyze the results by visualizing the primary concepts of interest both of the query and the retrieved museums, showing that there are nuances to the evaluation not captured by standard protocols. In particular, these nuances are more aligned to what a real user could search for using a retrieval tool, as the queries are likely short and focus on specific yet vaguely explained details which are interesting for the user. Therefore, there is a need for user studies to assess the performance of Text-to-Museum retrieval systems. Another future direction is related to how the museum descriptions, very long and detailed, may be very different to the user queries, short and vague. This reflects into the way text is processed, since we are relying on a complex pipeline made of sentence-level CLIP features and BiGRU processing, as only the museum descriptions are considered. Further investigation should be directed towards adding into the training process text data more similar to user queries.

7 Conclusions

As the Metaverse keeps growing, more users are facing difficulties in searching for experiences they find interesting, especially when looking for museums in the Metaverse. However, there are not intelligent tools supporting the users in the search process. In this paper, we started to investigate this novel task, Text-to-Museum retrieval, by collecting Museums3k, a large dataset of 3000 realistic museums, each containing many rooms and paintings. We implemented HierArtEx, a retrieval system leveraging hierarchical representations for modeling the museums, while using a mixture of generic and art-specific knowledge to model the visual contents of the rooms. We validated HierArtEx on Museums3k, observing that it achieves up to 47.7% R@1, and supported our design choices through ablation studies. We realized a qualitative analysis by leveraging concepts extracted from the queries and the retrieved museums, confirming the effectiveness of HierArtEx while highlighting some shortcomings in applying standard evaluation protocols for this complex problem. Finally, we highlighted some future research directions based on the findings of this work.

Acknowledgments. This work was supported by the PRIN 2022 "MUSMA" - CUP G53D23002930006 - "Funded by EU - Next-Generation EU - M4 C2 I1.1", and by the Department Strategic Plan (PSD) of the University of Udine – Interdepartmental Project on Artificial Intelligence (2020-25).

References

1. Abdari, A., Falcon, A., Serra, G.: Farmare: a furniture-aware multi-task methodology for recommending apartments based on the user interests. In: Proceedings of the IEEE/CVF International Conference on Computer Vision Workshops. pp. 4293–4303 (2023)
2. Abdari, A., Falcon, A., Serra, G.: Metaverse retrieval: Finding the best metaverse environment via language. In: Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval. pp. 1–9 (2023)
3. Abdari, A., Falcon, A., Serra, G.: Adoctera: Adaptive optimization constraints for improved text-guided retrieval of apartments. In: Proceedings of the ACM International Conference on Multimedia Retrieval. pp. 1043–1050 (2024)
4. Abdari, A., Falcon, A., Serra, G.: A language-based solution to enable metaverse retrieval. In: International Conference on Multimedia Modeling. pp. 477–488. Springer (2024)
5. Anastasovitis, E., Roumeliotis, M.: Transforming computed tomography scans into a full-immersive virtual museum for the antikythera mechanism. Digital Applications in Archaeology and Cultural Heritage **28**, e00259 (2023)
6. Barszcz, M., Dziedzic, K., Skublewska-Paszkowska, M., Powroznik, P.: 3d scanning digital models for virtual museums. Computer Animation and Virtual Worlds **34**(3-4), e2154 (2023)
7. Cooper, J.P., Wetherelt, A., Zazzaro, C., Eyre, M.: From boatyard to museum: 3d laser scanning and digital modelling of the qatar museums watercraft collection, doha, qatar. International journal of nautical archaeology **47**(2), 419–442 (2018)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. In: International Conference on Learning Representations (2020)
9. Garcia, N., Vogiatzis, G.: How to read paintings: semantic art understanding with multi-modal retrieval. In: Proceedings of the European Conference on Computer Vision (ECCV) Workshops. pp. 0–0 (2018)
10. Grincheva, N.: Cultural diplomacy under the "digital lockdown": Pandemic challenges and opportunities in museum diplomacy. Place Branding and Public Diplomacy **18**(1), 8 (2022)
11. Hegde, D., Valanarasu, J.M.J., Patel, V.: Clip goes 3d: Leveraging prompt tuning for language grounded 3d recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2028–2038 (2023)
12. ICOM: Report. museums, museum professionals and covid-19: a survey, https://icom.museum/wp-content/uploads/2021/07/Museums-and-Covid-19_third-ICOM-report.pdf [Accessed: August 17, 2024]
13. Karp, C.: Digital heritage in digital museums. Museum international **56**(1-2), 45–51 (2004)
14. Kiourt, C., Koutsoudis, A., Pavlidis, G.: Dynamus: A fully dynamic 3d virtual museum framework. Journal of Cultural Heritage **22**, 984–991 (2016)
15. Le, T.N., Nguyen, T.V., Le, M.Q., Nguyen, T.T., Huynh, V.T., Do, T.L., Le, K.D., Tran, M.K., Hoang-Xuan, N., Nguyen-Ho, T.L., et al.: Textanimar: text-based 3d animal fine-grained retrieval. Computers & Graphics **116**, 162–172 (2023)
16. Liu, M., Shi, R., Kuang, K., Zhu, Y., Li, X., Han, S., Cai, H., Porikli, F., Su, H.: Openshape: Scaling up 3d shape representation towards open-world understanding. Advances in neural information processing systems **36** (2024)

17. Merella, M., Farina, S., Scaglia, P., Caneve, G., Bernardini, G., Pieri, A., Collareta, A., Bianucci, G.: Structured-light 3d scanning as a tool for creating a digital collection of modern and fossil cetacean skeletons (natural history museum, university of pisa). Heritage **6**(10), 6762–6776 (2023)
18. Pei, X., Fu, S., Jiang, T.: An empirical study on user experience evaluation of vr interface in digital museums. Data and Information Management **7**(4), 100057 (2023)
19. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021)
20. Schroff, F., Kalenichenko, D., Philbin, J.: Facenet: A unified embedding for face recognition and clustering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 815–823 (2015)
21. Škola, F., Rizvić, S., Cozza, M., Barbieri, L., Bruno, F., Skarlatos, D., Liarokapis, F.: Virtual reality with 360-video storytelling in cultural heritage: Study of presence, engagement, and immersion. Sensors **20**(20), 5851 (2020)
22. Tan, W.R., Chan, C.S., Aguirre, H., Tanaka, K.: Improved artgan for conditional synthesis of natural image and artwork. IEEE Transactions on Image Processing **28**(1), 394–409 (2019). https://doi.org/10.1109/TIP.2018.2866698, https://doi.org/10.1109/TIP.2018.2866698
23. Xue, L., Gao, M., Xing, C., Martín-Martín, R., Wu, J., Xiong, C., Xu, R., Niebles, J.C., Savarese, S.: Ulip: Learning a unified representation of language, images, and point clouds for 3d understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 1179–1189 (2023)
24. Yu, F., Wang, Z., Li, D., Zhu, P., Liang, X., Wang, X., Okumura, M.: Towards cross-modal point cloud retrieval for indoor scenes. In: International Conference on Multimedia Modeling. pp. 89–102. Springer (2024)
25. Zhou, J., Wang, J., Ma, B., Liu, Y.S., Huang, T., Wang, X.: Uni3d: Exploring unified 3d representation at scale. arXiv preprint arXiv:2310.06773 (2023)
26. Zidianakis, E., Partarakis, N., Ntoa, S., Dimopoulos, A., Kopidaki, S., Ntagianta, A., Ntafotis, E., Xhako, A., Pervolarakis, Z., Kontaki, E., et al.: The invisible museum: A user-centric platform for creating virtual 3d exhibitions with vr support. Electronics **10**(3), 363 (2021)