HM3: Hierarchical Modeling of Multimedia Metaverses on 10000 Thematic Museums via Theme-aware Contrastive Loss Function

Gianluca Macrì
macri.gianluca@spes.uniud.it
Università degli Studi di Udine
Udine, Italy
Università degli Studi di Napoli Federico II
Napoli, Italy

Lorenzo Bazzana
bazzana.lorenzo@spes.uniud.it
Università degli Studi di Udine
Udine, Italy

Alex Falcon
falcon.alex@spes.uniud.it
Università degli Studi di Udine
Udine, Italy

Giuseppe Serra
giuseppe.serra@uniud.it
Università degli Studi di Udine
Udine, Italy

Abstract

The Metaverse and its immersive environments are gaining significant attention due to their potential applications across various fields, from healthcare to art. As their numbers grow, it becomes difficult to effectively search through them and identify those of interest to the user. Recently, Metaverses were modeled as multimedia-rich 3D scenarios. However, existing works on retrieving them via text have several shortcomings, including the lack of experimentation with joint analysis of heterogeneous multimedia formats within the Metaverse, the use of small-scale datasets with randomly aggregated elements, and the consequent lack of thematic coherence in retrieval methods. To address these issues, we introduce SAVAGE, a novel synthetic dataset of 10,000 thematic exhibitions containing both real-world paintings and generated video artworks. Moreover, we propose HM3, a new hierarchical methodology for Metaverse Retrieval which captures all the contents of the room and integrates both images and videos, while its training is guided by a novel theme-aware loss function. Experiments on SAVAGE demonstrate the effectiveness of HM3 in modelling museums. The method also shows considerable improvements on an existing dataset of Metaverses, with ablation studies and qualitative analyses confirming the utility of the proposed theme-aware loss function.

CCS Concepts

• **Information systems** → **Multimedia and multimodal retrieval**.

Keywords

Text-Museum Retrieval, Contrastive loss, Custom loss function, Multimedia, Artistic Metaverses, Metaverse Retrieval

ACM Reference Format:
Gianluca Macrì, Lorenzo Bazzana, Alex Falcon, and Giuseppe Serra. 2025. HM3: Hierarchical Modeling of Multimedia Metaverses on 10000 Thematic Museums via Theme-aware Contrastive Loss Function. In *Proceedings of the 2025 International Conference on Multimedia Retrieval (ICMR '25), June 30-July 3, 2025, Chicago, IL, USA.* ACM, New York, NY, USA, 9 pages. https://doi.org/10.1145/3731715.3733358

ICMR '25, Chicago, IL, USA

ACM ISBN 979-8-4007-1877-9/2025/06
https://doi.org/10.1145/3731715.3733358

1 Introduction

The metaverse, defined as an "immersive reality environment enabling ubiquitous access, identity, interoperability, and scalability" [12], is getting increasingly more attention over recent years, especially from investors and big techs like Microsoft and Meta. This is due to the vast variety of potential applications that it can offer in many different fields, ranging from healthcare with virtual training spaces for surgeons [28], to the art domain allowing artists to create and display their work digitally [21].

As noted in existing research [14], the number of these immersive 3D environments is expected to grow dramatically as the Metaverse becomes a mainstream medium. This evolution requires the development of retrieval systems that combine techniques from information retrieval and artificial intelligence to help users navigate and discover Metaverses that better align with their interests.

In this context, the problem of *Metaverse Retrieval* [4] was born, defined as the retrieval of complex 3D scenes which integrate multiple forms of multimedia content, such as images, videos, and 3D objects. Unlike traditional retrieval tasks, which treat all these visual elements *independently*, Metaverse Retrieval must consider them *jointly*, analyzing their relations and how they influence the relevance to user interests. For example, a virtual art exhibition might blend 2D artworks with interactive video guides or video artworks and three-dimensional sculptures to create a coherent experience. Jointly modeling these diverse formats is crucial both due to their potential complementarity of information and for enabling user queries that specify precise modalities—e.g. explicitly asking to include videos to guide observation or provide historical context.

Although the recency of the Metaverse Retrieval task, several works collected datasets in this direction and developed methodologies for it [1, 3, 4, 13]. However, we highlight three shortcomings, leaving a considerable gap with more realistic scenarios and real-world applications. First, prior research has analyzed separately the images [1, 13] and videos [4] found throughout the collected Metaverses, neglecting the more realistic scenario in which both types are available and important for the user. Second, existing

datasets have a reduced scale and aggregate elements randomly, neglecting that Metaverses usually have a central theme and are thus curated in that direction. For instance, in the context of digital museums, which represent a desirable application of the Metaverse Retrieval problem [13], the absence of a central theme is rather distant from the careful theme-based selection of art curators in real-world exhibitions. Third, as the presence of thematic coherence was not explored in the methods developed so far, it is not leveraged to support the retrieval system.

We address the shortcomings through three main contributions:

- We collect SAVAGE, a novel synthetic dataset of 10,000 artistic Metaverses, which marks an important step up from existing datasets, as the exhibitions are thematic and contain both real-world paintings and generated video artworks. This makes the elements of SAVAGE closer to expert-curated exhibitions, which is more realistic for both real-world exhibitions and existing digital museums.
- We propose a methodology, named HM3, that addresses the Metaverse Retrieval problem. It models artistic Metaverses through multiple steps: first, it captures the contents of the rooms, including images and videos; then, it progressively integrates them according to the hierarchical structure of museums. A vision-language embedding space is learned by optimizing two novel theme-aware loss functions, which enforce both inter-class and intra-class constraints.
- The experiments conducted on SAVAGE support the effectiveness of the proposed HM3 in modeling the museums, achieving 74.3% mAP and 94.9% nDCG. Moreover, such architecture is also tested on Museums3k [13], improving the R@1 by up to 36.5%. The ablation studies confirm the usefulness of the proposed double-loss strategy, leading to +33% mAP and +10% nDCG compared to using the standard triplet loss. Finally, qualitative analyses demonstrate the improved quality of the ranking lists via a better organization of the joint embedding space.

2 Related work

2.1 Metaverse Retrieval

Metaverse retrieval is a novel cross-modal task that emerged in the last few years regarding the problem of matching a textual query with a complex 3D scenario [4, 1]. Metaverses are characterized by their diversity and dynamism, and they are supposed to contain many elements that contribute to creating an immersive environment for the users. These elements may be 3D objects, like furniture and sculptures, or multimedia content like images and videos, and both play a crucial role in characterizing the Metaverses, e.g. in the case of digital museums [14]. Hence, it is fundamental to consider them when analyzing user queries to properly address their needs.

Such a scenario introduces two layers of additional difficulties compared to the simpler task of retrieving single 3D objects, which represents a longstanding research problem [9, 23, 32].

First, the presence of multiple 3D objects to be modeled at once, such as furniture in apartments. Recent works propose to represent the 3D space with multiple viewports to capture the elements forming the scene. [3] and [2] apply this approach to digital apartments, supporting the user's search during relocations. Specifically, [3] simultaneously trains the model for retrieval, while solving an image classification task regarding the furniture elements, forcing the model to consider the intricate internals. [2] introduces an adaptive distance constraint based on the similarity between apartments to improve the quality of the ranking. Conversely, [40] uses both appearance and depth information to model 3D scenes.

Second, the presence of additional multimedia elements that are meaningful to the user's interests. In [1], video elements were considered, including them in 3D furnished apartments through the careful positioning of digital screens capable of playing them. Other recent works focused on 2D multimedia elements in the form of paintings, which may be hanging on house walls [4] or be displayed inside some digital museums [13].

To investigate the retrieval of such complex scenes, a few annotated datasets have been proposed, mainly in the domain of furnished apartments and digital museums. For the former, both point cloud-based, like CRISP [40], and viewport-based datasets [3] are available. For the latter, which are the main focus of this work, the primary reference dataset was introduced in [13], in which the authors generated a synthetic collection of 3000 art-based museums.

Although this last dataset significantly improved the previously available options for artistic Metaverses, which used 3D apartments as the exhibition space and included a single painting per scene [1], it comes with a major shortcoming, as the selection of the artworks for each museum is random and not theme-based. We argue that this introduces a significant gap with real-world curated exhibitions, which are typically dedicated to a central theme or concept. Additionally, this dataset only deals with multimedia elements in the form of 2D paintings, without considering the video component.

Conversely, in [4], the dataset collected as a starting point for the problem of text-to-Metaverse retrieval solely consists of 3D scenarios based on multimedia elements in the form of videos. In this case, the limitation lies in the lack of images, usually central for museums, and in the presence of a single video element per scene. In [37], the authors explored several Metaverses and captured videos in each of them. However, the absence of textual descriptions and art-related Metaverses limits its suitability for our task.

In this work, we address these limitations by collecting a novel dataset (Sec. 3) of thematic museums with both 2D artworks and artistic videos.

2.2 Customizing the loss function to the task

Recent approaches for tackling cross-modal retrieval are based on the concept of learning a mapping of the input features to a joint embedding space, where similar elements are close. To reach this goal, contrastive loss functions have been successfully used for several years [6, 18]. Recently, they have become more popular, especially in the case of multimedia retrieval, thanks to advancements of models like MoCo [19], SimCLR [10], and CLIP [31].

One of the most common approaches consists of using an element as the anchor to which similar elements (positives) are pulled towards it, so that they are closer to it than dissimilar elements (negatives) by at least a fixed distance [34]. A similar concept was also applied to pairs [18], quadruplets [11], and larger groups [35].

Over time, researchers found that a fixed distance is often inadequate when intra-class variability and imbalance are present. This

resulted in methods employing dynamic, class-aware margins—such as using class prototypes to select positives and negatives relative to class centers—which also reduce the number of triplets and lower computational cost [30]. Other studies modify the margin itself, e.g. proposing an angular triplet loss that leverages cosine similarity between embeddings and their corresponding class centers [24], or introducing adaptive, class-aware margins [38], possibly linked to intra-class statistical properties [7]. Similar principles were also followed to decide the margin for video and 3D scene retrieval based on external constraints [2, 15]. More recently, dynamic margin strategies that evolve during training have been explored [39], along with frameworks to better manage class imbalance [8].

In this work, we introduce a novel loss function leveraging the central theme of the artistic Metaverses to apply different inter-class and intra-class constraints.

3 SAVAGE dataset

As mentioned before, existing datasets for the Metaverse Retrieval task either are devoid of multimedia elements, i.e. apartments [3] or rooms [40], or they include theme-agnostic exhibitions, i.e. non-artistic videos depicting single actions [4], single paintings [1], or randomly picked paintings [13]. Therefore, we collected the "**S**em**A**rt **V**ideo **A**rt **G**enerated **E**xhibitions" dataset consisting of 10,000 thematic artistic Metaverses, including multimedia elements in the form of both real-world paintings and generated artistic videos. With SAVAGE, the aim is to create a larger and more realistic dataset to foster research on the Metaverse Retrieval task and specifically for the Museum Retrieval task.

To overcome the shortcomings of existing datasets, which used pairs of decorated 3D scenes and textual descriptions [13], we define our dataset of N_E artistic Metaverses as $E = \{(s_i, d_i, t_i) : i \in 1 \dots N_E\}$, where each triplet associates a description d_i with both the 3D scene s_i and its theme t_i. The addition of this last element, absent in [13], is of central importance as real-world exhibitions are curated by experts, who generally design them based on a theme, e.g., a famous painter, an art school, or a specific timeframe [27].

A bottom-up approach has been followed to create our dataset, starting from collecting the single artworks (Sections 3.1 and 3.2), and then moving to the creation of the proper exhibitions (Sec. 3.3).

3.1 Image art data

As a first step, we selected the source for the paintings, looking for publicly available datasets [5, 16, 33, 36]. After a careful comparison, we selected SemArt, as it offers a reasonable amount of paintings (over 21k), each described by a human annotator and categorized under a wide range of metadata, such as "artist", "type", and "style". Specifically, the "type", which includes still-lives, portraits, and religious paintings, has a reasonable granularity. This allows us to obtain a good variability for the exhibitions within each theme. In the following, we refer to this dataset as $P = \{(p_i, d_i^P, t_i) : i \in 1 \dots N_P\}$, where p_i, d_i^P, and t_i correspond respectively to the image of an artwork, its description, and the type associated with it.

As SemArt includes an official three-fold split, we use the notation P_f where $f \in \{\text{train}, \text{val}, \text{test}\}$ to refer to the specific fold, while P is used to refer to any of the aforementioned elements when the operations are executed in the same way for each fold. The only difference we introduced to the original split is that we filtered out those elements that had a repeated value of d_i^P within the fold to avoid clashes between descriptions and different images.

3.2 Video art data

Contrary to the image case, selecting a dataset for video artworks posed a challenge. Despite the existence of a variety of large video datasets [20, 29], none of them is focused on art. After an initial analysis, even collecting a sufficient number of proper video artworks, i.e. videos intended to be artworks *per se*, appeared to be a challenge itself. First, the lack of *freely available* large-scale collections of digitalized videos regarding this topic. Second, even looking among existing collections, like UbuWeb, the videos are often associated with extremely noisy descriptions, if any. Additionally, it is difficult to pair last-century artistic videos with paintings from the Renaissance or even earlier time frames, making it nearly impossible to create coherent collections with the available themes.

Therefore, we opted for a different strategy, exploiting the images and their description to create 5s-long videos matching their content and themes through the use of a generative model. This way, we overcame the data availability issue, obtaining a large collection of elements automatically aligned with the artworks, albeit at the cost of moving a bit further away from real-world exhibitions, as the generative models may hallucinate and the results consist of short single-scene videos. Specifically, we used pyramid-flow [22], a novel text-to-video model that strikes a good compromise between inference speed and generation quality when compared with other alternatives like allegro [41] and AnimateDiff-Lightning [25].

To design the prompts for video generation, we processed the paintings' descriptions d_i^P through a large language model, so that the resulting videos could be directly related to the paintings. This step, depicted in Figure 1, has the primary objective of injecting into the prompt some dynamic action and camera movement, otherwise lacking in the static descriptions of the paintings, making the prompt more suitable for video generation. Specifically, we used Llama-3.2 3B-Instruct [17] to transform each description d_i^P into a textual prompt d_i^V that got used to generate a video v_i.

Through these steps, we were able to generate a dataset $V = \{(v_i, d_i^V, t_i) : i \in 1..N_P\}$ consisting of N_P triplets made by a video v_i, its corresponding textual description d_i^V, and the associated category t_i, derived from the corresponding i-th image artwork.

3.3 Exhibitions generation

As both P and V share a common categorization style, we can exploit them to generate the artistic Metaverses of the SAVAGE dataset. We used Unity to procedurally create the 3D scenes and decorate them with artworks of a specific theme. The museums are structured as an ordered series of 5 to 8 rooms, connected one to the following through a door and a short corridor (see Fig. 2). For each room, two of the four inner walls can display two artworks each, except for the final room, which has three walls available, allowing six artworks to be shown, and the first empty one. A similar approach was followed to create Museums3k [13], except that videos and themes were not considered there.

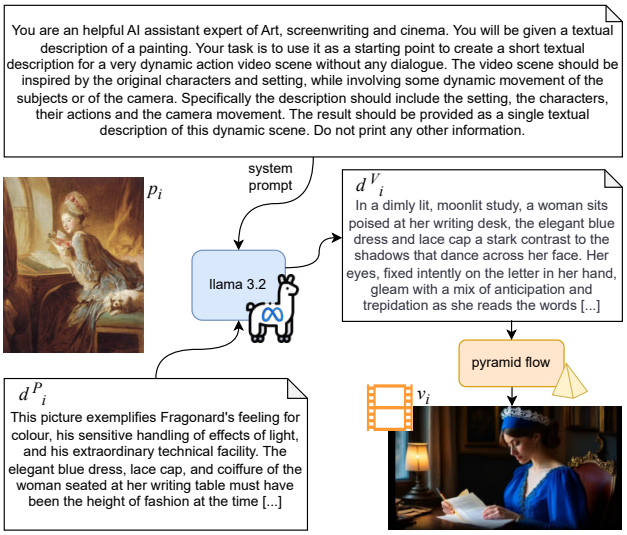

Figure 1: A scheme of the video generation pipeline. For each painting p_i, the matching description d^P_i is processed through llama3.2 along with a custom system prompt to create d^V_i, then used as a prompt by a text-to-video model, pyramidflow, to generate the video v_i.

To generate the full dataset E, we followed a three-way subdivision into test, validation, and test, with 7K, 1K, and 2K elements, respectively. To generate a pair (s_i, d_i) for the fold f, the first step consisted in sampling the theme among those in the set $\{t : |\{(p_j, d_j, t) : (p_j, d_j, t) \in P_f\}| \geq 30 \wedge t \neq \text{“other”}\}$ with a probability proportional to its numerosity. This means that, for each fold f, we considered themes t with enough elements of each modality to fully decorate our largest possible exhibition, excluding the “other” theme. This guarantees that no artworks are repeated within any exhibition. Once t has been fixed, the number of rooms is randomly selected, and for each displaying location, a specific artwork with theme t is sampled from $P \cup V$.

Finally, a description is associated with each scene to complete E. Following existing approaches [3, 13], each d_i is created rather verbosely following a predefined template. This includes an introductory sentence about the structure of the exhibition, followed by a series of sentences describing each of the artworks displayed in the same order a visitor would encounter them (Figure 2).

4 Proposed methodology: HM3

As the artistic Metaverses naturally follow a hierarchical structure, we modeled them leveraging this organization. The proposed architecture HM3, depicted in Figure 3, is organized into three main modules dedicated to processing the exhibition containing both paintings and art videos (Section 4.1), the textual description (Sec. 4.2), and to learn the joint museum-description embedding space through a novel theme-aware loss function (Sec. 4.3). To make the discussion clearer, we will drop the subscript that refers to the specific instance of the scene or the video, using simply s and d.

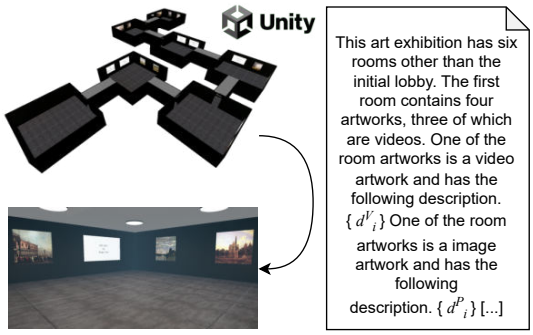

Figure 2: Example of a procedurally generated museum s_i (upper left), a view of its decorated internals (lower left), and the template for the matching textual description (right).

4.1 Art exhibition processing

Leveraging a hierarchical structure to process the exhibition entails processing the single rooms and their contents, and then aggregating everything into a museum-level representation. For modeling the room, a virtual camera is placed in its center and rotated by 90 degrees to take 4 screenshots of the distinct inner walls, similarly to [13]. However, our exhibitions also include videos, whose contents cannot be naïvely captured by such a strategy, which could also introduce ambiguity as a still frame of a video may be confused with a 2D artwork, and vice versa. To allow the model to properly process images and videos, the latter are visualized with a default screen displaying their title and author on a plain white background when using the virtual camera. Then, to include the actual video information, we assume to be able to extract the videos so that they can be fed directly to our model. To satisfy such an assumption in a real-world scenario, the model would need to either access this information through the metadata or to individuate the video artworks and record them. Each video, represented by 32 sampled frames, is processed through the pre-trained visual encoder of CLIP4Clip [26], whereas a pre-trained version of CLIP [31] encodes the screenshots. In both cases, the ViT-B/32 version was used.

Consequently, for a specific room j of the Metaverse s, we obtain a feature matrix $x_{pov} \in \mathbb{R}^{4 \times D_{pov}}$ representing the 4 internal screenshots, and $x_{vid} \in \mathbb{R}^{N^{s,j}_{vid} \times D_{vid}}$ where $N^{s,j}_{vid}$ is the number of video elements of room j. These matrices are then aggregated through f_{room} to create a single representation for the room, i.e. $f_{room}(x_{pov}, x_{vid}) = x_{room} \in \mathbb{R}^{D_{room}}$. Specifically, each screenshot represented by a row of x_{pov} is transformed through a linear layer followed by a ReLU, and the average of the transformed vectors is computed to obtain an aggregated representation of the screenshots $x'_{pov} \in \mathbb{R}^{D'_{pov}}$. Analogously, x_{vid} undergoes a similar transformation with a separate linear layer followed by a ReLU to obtain $x'_{vid} \in \mathbb{R}^{D'_{vid}}$. These two are finally concatenated in $\mathbb{R}^{D'_{vid}+D'_{pov}}$ and processed through a new linear layer to obtain a x_{room}.

The room features $x^j_{room}, j \in \{1 \dots N^s_{room}\}$ are then aggregated to obtain a single representation $x_i \in \mathbb{R}^D$ through a function f_{exhib}. To implement it, we use a bidirectional GRU with a hidden layer of size D and then take the average of the two final hidden states. Note that this decision forces an order on the museum visit.

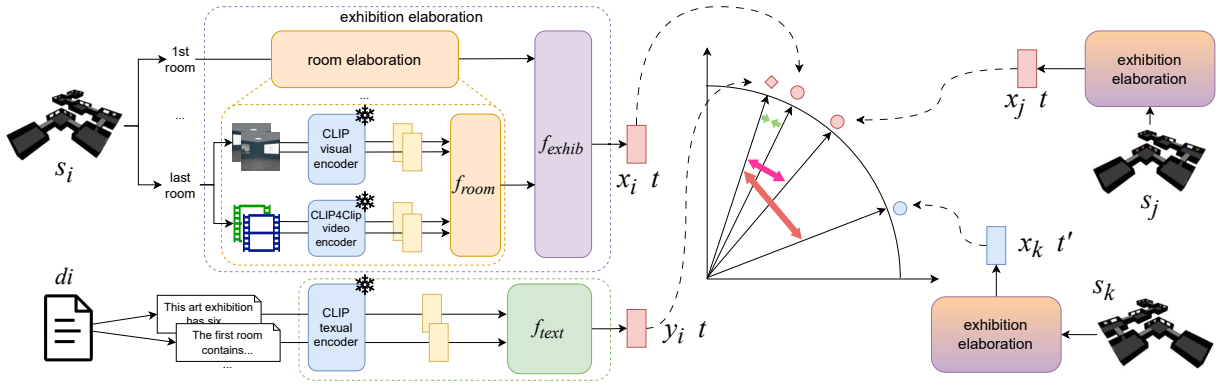

Figure 3: Overview of the proposed methodology, HM3. The museum s_i is processed room by room, using the appropriate encoders for its screenshot and video and aggregating the information with f_{room}. Then, f_{exhib} joins the embeddings of the rooms into a museum-level representation. The sentences of description d_i are encoded and then aggregated by f_{text}. Finally, the theme-aware two-components loss function has two aims: L_{DTN} favours (green arrows) the alignment of a museum s_i and its description d_i while pushing away (orange arrows) s_k as they have different themes (t and t'); while L_{STN} aims at preserving a little distance between the non-matching elements within the theme t (magenta arrows).

4.2 Textual description processing

As the descriptions d_i are rather verbose and the context window of the textual encoder is rather limited, we split them on a sentence-by-sentence level before extracting the corresponding initial features $y_i^j \in \mathbb{R}^{D_{sent}}$. The sequence of these is then further processed and aggregated through a function f_{text}, resulting in a final vector $y_j \in \mathbb{R}^D$ that represents the whole description. To implement f_{text}, we first use a bi-GRU with a hidden layer of size D, and then take the average of the two final hidden states, similarly to [3, 13].

4.3 A customized theme-aware loss function

The objective of our training is to align the vectorial representations x_i and y_i output by f_{exhib} and f_{text} when the corresponding museum and description are *similar* while distancing them from dissimilar elements. When learning a cross-modal retrieval system, it is common to consider x_i and y_i similar only when they form a pair in the dataset (e.g. an image and its description) [31]. In our scenario, given our dataset $E = \{(s_i, d_i, t_i) : i \in 1..N_E\}$, we consider two elements indexed i and j to be similar in two cases. The first, similar to the standard approach followed in cross-modal literature, is the exact matching, i.e. the elements form a pair in the E. The second case arises from the presence of a common theme. In this scenario, we would like to have x_i, y_i, x_j, and y_j more similar to each other with respect to x_k and y_k, when $t_i = t_j \neq t_k$. This is because if exhibitions i and j share the theme, e.g. both contain religious artworks, then they are likely more similar to each other than the exhibition k which follows a different theme (e.g. still-lives).

To achieve these two goals, we propose to use a novel two-component theme-aware loss function $\mathcal{L}$ which leverages the *theme* as a key innovative element. The first component, L_{DTN}—short for "different theme negative"—acts similarly to the standard all negative triplet loss [34], i.e. aligning a pair (x_i, y_i) while enforcing a margin with a "negative" element x_j, but this is picked only among those elements belonging to a different theme $t_j \neq t_i$. For a batch $B \subseteq E$ and a similarity metric $sim(.)$, L_{DTN} is defined as follows:

$$\mathcal{L}_{DTN} = \sum_{\substack{(s_i,d_i,t_i),\\ (s_j,d_j,t_j)\in B,\ t_i\neq t_j}} \frac{L_{\Delta_{DTN}}(x_i, y_i, y_j) + L_{\Delta_{DTN}}(y_i, x_i, x_j)}{2\overline{M}}$$

where $\overline{M} = |\{(i,j) : (s_i, d_i, t_i), (s_j, d_j, t_j) \in B, t_i \neq t_j\}|$, and $L_\Delta(a,p,n) = \max(0, sim(a,n) - sim(a,p) + \Delta)$.

While this first component avoids enforcing a dissimilarity with elements of the same theme, it does not penalize the case in which a different element s_j is more similar to d_i than s_i. Since a stronger alignment between the features corresponding to an exactly matching pair is desirable, we define the second component L_{STN}, short for "same theme negative", as follows:

$$\mathcal{L}_{STN} = \sum_{\substack{(s_i,d_i,t_i),\\ (s_j,d_j,t_j)\in B,\ t_i=t_j,\ i\neq j}} \frac{L_{\Delta_{STN}}(x_i, y_i, y_j) + L_{\Delta_{STN}}(y_i, x_i, x_j)}{2M}$$

where $M = |\{(i,j) : (s_i, d_i, t_i), (s_j, d_j, t_j) \in B \wedge t_i = t_j \wedge i \neq j|\}$. Note that while it resembles the standard triplet loss, L_{STN} solely considers elements within the batch *sharing a common theme*. Additionally, as elements that share a common theme are still expected to be more similar than the others, we require $\Delta_{DTN} > \Delta_{STN}$.

Finally, $\mathcal{L}$ can be defined as a convex combination:

$$\mathcal{L} = (1-\alpha)\mathcal{L}_{DTN} + \alpha\mathcal{L}_{STN}$$

for a weight $\alpha \in [0,1]$. As side note, observe that when $\alpha = 0.5$ and $\Delta_{DTN} = \Delta_{STN}$, $\mathcal{L}$ collapses to the standard triplet loss.

5 Experimental results

5.1 Implementation details

To perform the experiments, we use the dataset E (Sec. 3), already split into train, validation, and test. For each configuration, we train the model for 25 epochs on E_{train} using a batch size of 64 while monitoring the performance on E_{val} to prevent overfitting. We use

Adam with learning rate 5e-4 and a step scheduler ($\gamma = 0.75$) applied every 5 epochs. To account for variability due to, e.g., random initialization of learnable layers, the results on E_{test} are reported as the mean text-to-Metaverse retrieval performance over three independent runs using the model obtained at the final epoch.

All the configurations have an initial feature size of $512 = D_{sent} = D_{pov} = D_{vid}$ and we use an output feature size $D = 256$. Additionally, for HM3 we set $D'_{vid} = D'_{vid} = 256$ and $D_{room} = 256$. For the loss we set $\Delta_{DTN} = 0.25$, $\Delta_{STN} = 0.15$ and $\alpha = 0.05$.

We conducted experiments on a server with an NVIDIA A100 GPU and an AMD EPYC 7643 (2.35GHz) processor. Our setup used Python 3.11.5, PyTorch 2.1.0, and CUDA 11.8. Both code and data are fully available at: https://github.com/gianlucamacri/HM3-ICMR25.

5.2 Baselines under analysis

b MP: A baseline not taking advantage of the hierarchical structure of the exhibitions. It is a variation of HM3 where we removed f_{exhib} entirely. The final representation for the museum $x_i \in \mathbb{R}^D$ is obtained by processing the x_{pov} and x_{vid} feature matrices for the whole exhibition directly through f_{room}, as in [13].

h MP + 1dCNN: An alternative version of HM3 where f_{exhib} is implemented using a 1D convolutional deep neural network. In this case, f_{exhib} processes the room features $x^j_{room}, j \in \{1 \dots N^s_{room}\}$ for an exhibition s_i through a kernel of size 3 and "same" padding to get intermediate representations with in $\mathbb{R}^{256}$. The average is then taken and transformed through a linear layer to get x_i.

b 1dCNN: An alterative version of our baseline that was used in [3, 2]. In this case, only f_{room} is used and it is implemented using a 1D convolutional deep neural network with a kernel of size 3 and "same" padding. This processing step is followed separately for x_{pov} and x_{vid}, which then get averaged, concatenated, and finally processed through a final linear layer.

5.3 Evaluation Metrics

To assess the performance, we use common evaluation metrics for retrieval, all based on the concept of relevance between a query and a target. As stated in Section 4.3, we have two ways to state when two elements are similar: either the exact matching provided in the dataset or, more broadly, the theme commonality.

This difference is encoded in the definition of a relevance function $r^q : \{1..N_E\} \to \{0, 1\}$ for a query element with index q. For the concept of exact matching, we use $r^q_{exact}(i)$ that takes the value 1 if $idx(i) = q$, where $idx(i)$ is the index of the i-th element in the ranking output, and 0 otherwise. Differently, for theme-based matching, we use $r^q_{theme}(i)$ that takes the value 1 if and only if $t_{idx(i)} = t_q$. In the following, we use r to indicate either of the two relevance functions, and use Q to denote the corresponding number of relevant elements for q, unless otherwise stated.

Based on these definitions, we consider the following metrics: Recall (**R@k**) at rank k for $k \in \{1, 5, 10\}$, median rank (**MedR**), mean average precision (**mAP**), and normalized discounted cumulative gain (**nDCG**).

The recall is defined as $R@k = \sum_{i=1}^{k} r(i)/Q$. Note that it may be that $M >> k$ when themes are considered in r, making this metric interesting only in the exact matching case.

To consider the position of the relevant elements within the ranking portion that gets analyzed, we also include metrics that look more broadly at the quality of the ranking: mAP and nDCG. The former is defined as the mean of the Average Precisions (AP) computed for various q, where $AP = \frac{1}{Q} \sum_{k=1}^{N_E} P@k \cdot r(k)$. While in our case the relevance is binary and not continuous, as is common for the nDCG, we included this metric to allow a comparison with future works that may rely on a finer-graded relevance. The nDCG is defined as $nDCG = \frac{DCG}{IDCG}$, where $DCG = \sum_{i=1}^{N_E} \frac{r(i)}{\log_2(i+1)}$ and the denominator is the Ideal Discounted Cumulative Gain (IDCG), computed in our case as $IDCG = \sum_{i=1}^{Q} \log_2^{-1}(i+1)$ which considers the optimal ranking to normalize for different values of Q.

5.4 Ablation studies and hyperparameters search

Weighting the loss components. The first experiment concerns the choice of the weights assigned to the two loss components. Figure 4 shows how the mAP, both when using r_{exact} and r_{theme}, varies for different values of α. This is computed for HM3 and the three baselines, with $\Delta_{DTN} = 0.25$ and $\Delta_{STN} = 0.15$. The figure clearly shows the contrast between the performance computed through the exact matching and that computed by theme. In particular, using a high α leads to high "exact matching" mAP (above 80%, with the exception of "h MP + 1dCNN") and low "theme-aware" mAP (between 30 and 50%). Conversely, a low α leads nearly 100% "theme-aware" mAP for every method, and less than 10% "exact matching" mAP. Notably, HM3 always represents a better compromise. Finally, we choose $\alpha = 0.05$ as a sweet plot value.

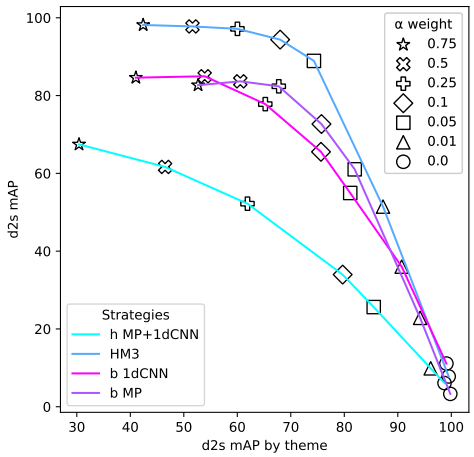

Figure 4: Comparison of HM3 and different baselines for the values of mAP in the case of exact matching and theme-based matching at different values of α.

Finding the value for Δ_{STN}. As L_{STN} is an important addition to the training function, we analyzed the response of the networks as Δ_{STN} varies. Δ_{DTN} was set to 0.25 and $\alpha = 0.05$. The results, reported in Table 1, align with the intuition that forcing a slightly larger margin (e.g. 0.20) within the elements that share a common theme leads to improvements for the exact matching metrics at the

cost of some loss when computed by theme. For instance, for HM3, going from $\Delta_{STN} = 0.15$ to 0.20, the metrics vary as follows: (exact matching) +5.5% mAP and +1.9% nDCG, (by theme) -5.7% mAP and -0.3% nDCG. The opposite behaviour can be seen when the margin is reduced instead. Overall, we chose $\Delta_{STN} = 0.15$ as it strikes a good balance when considering the average values.

Table 1: Performance comparison across both exact matching and theme-aware metrics as the margin for L_{STN} varies.

	Δ_{STN}	mAP	nDCG	mAP by theme	nDCG by theme	avg
HM3	0.10	78.4	83.5	**83.5**	**96.9**	85.6
	0.15	88.9	91.6	74.3	94.9	**87.4**
	0.20	**94.4**	**95.8**	65.4	92.7	87.1
h MP+1dCNN	0.10	19.3	34.8	**91.8**	**98.5**	61.1
	0.15	25.6	40.6	85.5	97.2	**62.2**
	0.20	**27.6**	**42.5**	79.8	96.0	61.5

Ablating the loss function. As the proposed loss function relies on two loss components, both of which have novelty, we analyze their performance both jointly and separately compared to the standard triplet loss (STL). We fixed $\alpha = 0.05$, $\Delta_{DTN} = 0.25$, and $\Delta_{STN} = 0.15$ for our loss, and $\Delta = 0.25$ for the triplet loss as typically done in existing works [13, 3]. As shown in Table 2, using the standard loss leads to better results for the exact matching, while sacrificing the theme-based metrics. Interestingly, using just L_{DTN} leads to excellent semantic match metrics (mostly above 99%), with very poor exact matching metrics. This is likely due to L_{DTN} focusing solely on pushing museums with different themes away while letting *all* the exhibitions with the same theme naturally form tight, yet internally messy, clusters. By adding L_{STN} and thus adding control within the clusters, both HM3 and "h MP + 1dCNN" reach a good balance of the two sets of metrics, with an average of 87.4% and 62.2% (+6.2% and +1.7% over STL, respectively).

Table 2: Comparison of HM3 and the baselines trained with either the proposed two-component loss ($\alpha = 0.05$, $\Delta_{DTN} = 0.25$, $\Delta_{STN} = 0.15$), part of it or the standard triplet loss (STL).

		Exact match		Semantic match		
	Loss	mAP	nDCG	mAP	nDCG	avg
HM3	DTN+STN	88.9	91.6	74.3	94.9	**87.4**
	DTN	7.8	22.1	**99.5**	**99.9**	57.3
	STL	**99.3**	**99.5**	41.2	84.6	81.2
h MP+1dCNN	DTN+STN	25.6	40.6	85.5	97.2	**62.2**
	DTN	6.1	20.9	**98.7**	**99.7**	56.4
	STL	**44.6**	**56.8**	52.7	88.0	60.5

Removal of video features. In the experiments so far, we used both x_{pov} and x_{vid} as inputs. It is important to ensure that introducing the videos is indeed beneficial as this would both mean that they contribute to a better retrieval system and that the model is learning to use them effectively. The results, reported in Table 3, confirm that in all cases there is a considerable improvement (up to 32%) under all metrics, with reduced impact on theme-based ones.

Table 3: Performance comparison across exact matching and theme-aware metrics as the methods consider only x_{pov} or also x_{vid}. Quantitative results support the usefulness of modeling both types to fully capture the Metaverse contents.

		Exact match		Semantic match	
		mAP	nDCG	mAP	nDCG
HM3	pov	77.7	83.0	68.8	93.0
	+video	**88.9**	**91.6**	**74.3**	**94.9**
h MP+1dCNN	pov	12.5	28.5	83.6	96.0
	+video	**25.6**	**40.6**	**85.5**	**97.2**
b 1dCNN	pov	36.7	50.3	78.6	95.2
	+video	**55.0**	**65.2**	**81.1**	**96.4**
b MP	pov	29.0	43.4	77.2	95.3
	+video	**61.0**	**69.7**	**81.9**	**96.6**

5.5 Qualitative analysis

Analyzing the learned embedding space. In Fig. 5, we use tSNE to analyze the embedding space learned by HM3 and compare it with the case using STL. The figure is aligned with the observations in Table 2. In fact, the proposed loss function (Fig. 5 left) imposes a stronger separation between thematic clusters of artistic Metaverses at the cost of losing some alignment between an exhibition and the corresponding description. Symmetrically, the model using the STL (Fig. 5 right) offers much less discrimination between different themes, while supporting slightly better exact matchings.

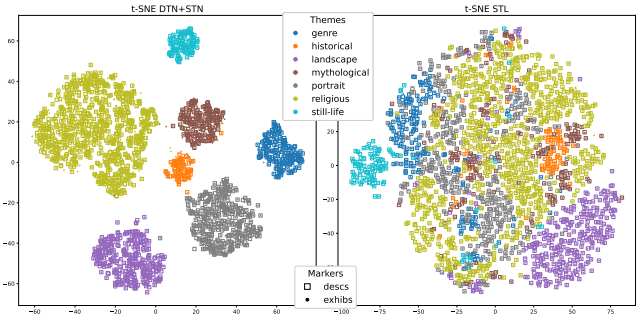

Figure 5: tSNE visualization of the joint exhibition-description embedding space defined by our hierarchical model when using the proposed two-component loss (left) and the standard triplet loss (right).

Plotting the ranking lists. To further highlight the importance of considering relevant also exhibitions that share a theme, we selected three test queries according to the quartiles of the distribution of thematic mAP for HM3. This way, cases where the model performs poorly, decently, or well are all represented. The results are visualized in Fig. 6 with three colors: cyan for the exact match, yellow for other relevant exhibitions, and black otherwise. We used HM3 and "b MP", both with the proposed loss and the STL. Overall, the four models generally retrieve the exact match

among the first ranks. However, the distribution of other relevant elements is wildly different, especially when the STL is used. The third query is representative of this issue, as the two models using the proposed loss are retrieving most of the relevant Metaverses early, as shown by the large yellow bands on the left, while they are scattered across the whole ranking list with the STL.

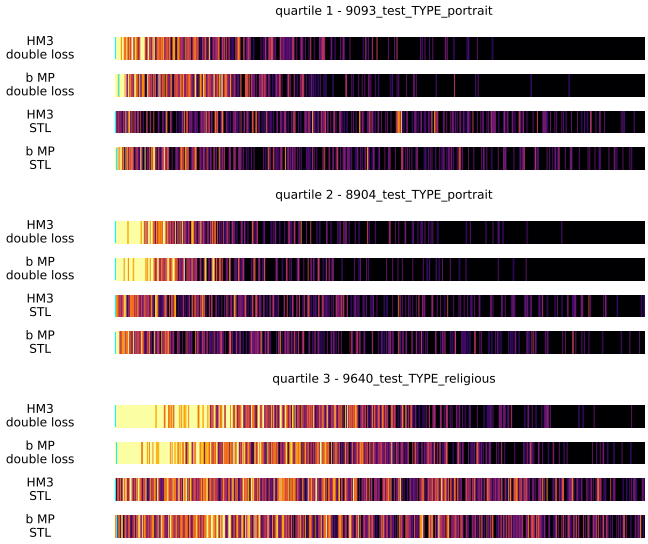

Figure 6: Visualization of the final ranking obtained by HM3 and "b MP" when using the double loss vs the standard triplet loss for different queries picked based on the quartiles of the thematic mAP of HM3. The exact match has a cyan colour, the same class elements use yellow bars, the others are black.

5.6 Comparison with other datasets

To further validate HM3, we tested it against the Museums3k dataset [13] using the standard triplet loss as no themes are available in this case. The same training data and experimental setting were used, relying on their open-source code to ensure a fair comparison. We first reproduced their models in our codebase, obtaining similar results with differences of less than 3%, likely due to stochastic aspects. We also trained a baseline which could be obtained by removing the function f_{room} and setting $D_{room} = 512$. The results, reported in Table 4, show that this baseline obtains comparable results (33.7% vs 36.6%) to their hierarchical approach without the additional art experts. However, the introduction of the bi-GRU, which forces an ordering on the rooms compared to HierArt[13], leads to considerable improvements, achieving 84.2% R@1.

6 Discussion and future works

While HM3 achieves strong results for both exact and thematic matching, we acknowledge a few important limitations. Regarding SAVAGE, as noted in Section 3.2, the generated videos cannot be equated with actual video artworks. Moreover, generative models are known to introduce hallucinations, increasing dissimilarity from real-world cases and creating gaps between textual descriptions and visual content. Additionally, as noted by [13], Metaverse descriptions are much longer and more detailed than typical user queries, often shorter and more ambiguous. Future work should collaborate with real-world art collections and conduct user studies to better model query structures, bridging the gap between research settings and real-world text-to-Metaverse retrieval scenarios. Moreover, while we focused on paintings and videos, future works should also integrate sculptures and interactive experiences within the scenarios, as these are commonly found in real-world museums and further enrich them.

Regarding our approach to Metaverse modeling, we recognize two key limitations. First, structuring exhibitions as a linear sequence of rooms is an oversimplification of both virtual and real-world exhibition spaces. Second, using a limited set of themes as a proxy for defining the exhibition similarity is a major simplification that may not fully align with user needs. Future work should explore more flexible spatial structures and, consequently, suitable models like Graph Neural Networks, as well as richer similarity metrics to improve retrieval effectiveness.

Table 4: Comparison of HM3 with the methods proposed in Museums3k [13] using the STL. Details in Sec. 5.6.

architecture	R@1	R@5	R@10	medR
HM3	**84.2** (2.0)	**97.3** (0.6)	**99.0** (0.3)	**1.0** (0.0)
only Bi-GRU	33.7 (4.3)	61.9 (4.2)	73.3 (2.7)	3.3 (1.2)
HierArtEx[13]	47.7	81.7	90.8	2.0
w/o ArtExp[13]	36.6	70.9	82.1	2.3
baseline[13]	14.2	40.4	56.7	8.7

7 Conclusions

In this work, we addressed some key limitations in text-to-Metaverse retrieval by introducing SAVAGE, a large-scale synthetic dataset of 10,000 thematic museums integrating both real-world paintings and generated video artworks, getting closer to the curated nature of real-world digital museums. To address the retrieval task while taking advantage of the thematic nature of the Metaverses, we proposed HM3, a new hierarchical approach that uses a novel two-component theme-aware loss function to learn a joint embedding space between exhibitions comprising multimedia elements in the form of images and videos, and corresponding textual descriptions. We validated our approach on SAVAGE, achieving 74.3% mAP and 94.9% nDCG, and supported our design choices through ablation studies and qualitative analysis. Furthermore, we significantly improved the previous state-of-the-art results for the Museums3k [13] reference dataset. Finally, we underlined the limitations of our work, suggesting some future research directions.

Acknowledgments

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

References

[1] Ali Abdari, Alex Falcon, and Giuseppe Serra. 2024. A language-based solution to enable metaverse retrieval. In *International Conference on Multimedia Modeling*. Springer, 477–488.

[2] Ali Abdari, Alex Falcon, and Giuseppe Serra. 2024. Adoctera: adaptive optimization constraints for improved text-guided retrieval of apartments. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, 1043–1050.

[3] Ali Abdari, Alex Falcon, and Giuseppe Serra. 2023. 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*, 4293–4303.

[4] Ali Abdari, Alex Falcon, and Giuseppe Serra. 2023. Metaverse retrieval: finding the best metaverse environment via language. In *Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval*, 1–9.

[5] Panos Achlioptas, Maks Ovsjanikov, Kilichbek Haydarov, Mohamed Elhoseiny, and Leonidas Guibas. 2021. Artemis: affective language for visual art. *CoRR*, abs/2101.07396.

[6] Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. 1993. Signature verification using a" siamese" time delay neural network. *Advances in neural information processing systems*, 6.

[7] Yannick Le Cacheux, Herve Le Borgne, and Michel Crucianu. 2019. Modeling inter and intra-class relations in the triplet loss for zero-shot learning. In *Proceedings of the IEEE/CVF international conference on computer vision*, 10333–10342.

[8] Kanghao Chen, Weixian Lei, Shen Zhao, Wei-Shi Zheng, and Ruixuan Wang. 2023. Pcct: progressive class-center triplet loss for imbalanced medical image classification. *IEEE Journal of Biomedical and Health Informatics*, 27, 4, 2026–2036. doi:10.1109/JBHI.2023.3240136.

[9] Kevin Chen, Christopher B Choy, Manolis Savva, Angel X Chang, Thomas Funkhouser, and Silvio Savarese. 2019. Text2shape: generating shapes from natural language by learning joint embeddings. In *Computer Vision–ACCV 2018: 14th Asian Conference on Computer Vision, Perth, Australia, December 2–6, 2018, Revised Selected Papers, Part III 14*. Springer, 100–116.

[10] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*. PmLR, 1597–1607.

[11] Weihua Chen, Xiaotang Chen, Jianguo Zhang, and Kaiqi Huang. 2017. Beyond triplet loss: a deep quadruplet network for person re-identification. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 1320–1329. doi:10.1109/CVPR.2017.145.

[12] John David N. Dionisio, William G. Burns III, and Richard Gilbert. 2013. 3d virtual worlds and the metaverse: current status and future possibilities. *ACM Comput. Surv.*, 45, 3, Article 34, (July 2013), 38 pages. doi:10.1145/2480741.2480751.

[13] Alex Falcon, Ali Abdari, and Giuseppe Serra. 2024. Hierartex: hierarchical representations and art experts supporting the retrieval of museums in the metaverse. In *International Conference on Multimedia Modeling*. Springer, 60–73.

[14] Alex Falcon, Beatrice Portelli, Ali Abdari, Giuseppe Serra, et al. 2024. Paving the way for personalized museums tours in the metaverse. (2024).

[15] Alex Falcon, Swathikiran Sudhakaran, Giuseppe Serra, Sergio Escalera, and Oswald Lanz. 2022. Relevance-based margin for contrastively-trained video retrieval models. In *Proceedings of the 2022 international conference on multimedia retrieval*, 146–157.

[16] Noa Garcia and George Vogiatzis. 2018. How to read paintings: semantic art understanding with multi-modal retrieval. In *Proceedings of the European Conference in Computer Vision Workshops*.

[17] Aaron Grattafiori et al. 2024. The llama 3 herd of models. (2024). https://arxiv.org/abs/2407.21783 arXiv: 2407.21783 [cs.AI].

[18] R. Hadsell, S. Chopra, and Y. LeCun. 2006. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. Vol. 2, 1735–1742. doi:10.1109/CVPR.2006.100.

[19] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross Girshick. 2020. Momentum contrast for unsupervised visual representation learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 9729–9738.

[20] Shawn Hershey et al. 2017. Cnn architectures for large-scale audio classification. In *2017 ieee international conference on acoustics, speech and signal processing (icassp)*. IEEE, 131–135.

[21] William Hurst, Orestis Spyrou, Bedir Tekinerdogan, and Caspar Krampe. 2023. Digital art and the metaverse: benefits and challenges. *Future Internet*, 15, 6, 188.

[22] Yang Jin et al. 2024. Pyramidal flow matching for efficient video generative modeling. *arXiv preprint arXiv:2410.05954*.

[23] Trung-Nghia Le et al. 2023. Textanimar: text-based 3d animal fine-grained retrieval. *Computers & Graphics*, 116, 162–172.

[24] Zhaoqun Li, Cheng Xu, and Biao Leng. 2019. Angular triplet-center loss for multi-view 3d shape retrieval. In *Proceedings of the AAAI conference on artificial intelligence* number 01. Vol. 33, 8682–8689.

[25] Shanchuan Lin and Xiao Yang. 2024. Animatediff-lightning: cross-model diffusion distillation. (2024). https://arxiv.org/abs/2403.12706 arXiv: 2403.12706 [cs.CV].

[26] Huaishao Luo, Lei Ji, Ming Zhong, Yang Chen, Wen Lei, Nan Duan, and Tianrui Li. 2022. Clip4clip: an empirical study of clip for end to end video clip retrieval and captioning. *Neurocomputing*, 508, 293–304. doi:https://doi.org/10.1016/j.neucom.2022.07.028.

[27] Freda Matassa. 2014. *Organizing Exhibitions: a handbook for museums, libraries and archives*. facet publishing.

[28] Kabir Matwala, Taner Shakir, Chetan Bhan, and Manish Chand. 2024. The surgical metaverse. *Cirugía Española*, 102, S61–S65.

[29] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, 2630–2640.

[30] Zuheng Ming, Joseph Chazalon, Muhammad Muzzamil Luqman, Muriel Visani, and Jean-Christophe Burie. 2017. Simple triplet loss based on intra/inter-class metric learning for face verification. In *2017 IEEE International Conference on Computer Vision Workshops (ICCVW)*, 1656–1664. doi:10.1109/ICCVW.2017.194.

[31] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmLR, 8748–8763.

[32] Yue Ruan, Han-Hung Lee, Yiming Zhang, Ke Zhang, and Angel X Chang. 2024. Tricolo: trimodal contrastive loss for text to shape retrieval. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, 5815–5825.

[33] Babak Saleh and Ahmed Elgammal. 2016. Large-scale classification of fine-art paintings: learning the right metric on the right feature. *International Journal for Digital Art History*, 2, (Oct. 2016). doi:10.11588/dah.2016.2.23376.

[34] Florian Schroff, Dmitry Kalenichenko, and James Philbin. 2015. Facenet: a unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823.

[35] Kihyuk Sohn. 2016. Improved deep metric learning with multi-class n-pair loss objective. In *Advances in Neural Information Processing Systems*. D. Lee, M. Sugiyama, U. Luxburg, I. Guyon, and R. Garnett, (Eds.) Vol. 29. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2016/file/6b180037abbebea991d8b1232f8a8ca9-Paper.pdf.

[36] Matteo Stefanini, Marcella Cornia, Lorenzo Baraldi, Massimiliano Corsini, and Rita Cucchiara. 2019. Artpedia: A New Visual-Semantic Dataset with Visual and Contextual Sentences. In *Proceedings of the International Conference on Image Analysis and Processing*.

[37] Patrick Steinert, Stefan Wagenpfeil, Ingo Frommholz, and Matthias L Hemmje. 2024. 256 metaverse records dataset. In *Proceedings of the 32nd ACM International Conference on Multimedia*, 4256–4263.

[38] Yi Tian, Zhiwei Wen, Weicheng Xie, Xi Zhang, Linlin Shen, and Jinming Duan. 2019. Outlier-suppressed triplet loss with adaptive class-aware margins for facial expression recognition. *2019 IEEE International Conference on Image Processing (ICIP)*, 46–50. https://api.semanticscholar.org/CorpusID:202786173.

[39] Weicheng Xie, Haoqian Wu, Yi Tian, Mengchao Bai, and Linlin Shen. 2022. Triplet loss with multistage outlier suppression and class-pair margins for facial expression recognition. *IEEE Transactions on Circuits and Systems for Video Technology*, 32, 2, 690–703. doi:10.1109/TCSVT.2021.3063052.

[40] Fuyang Yu, Zhen Wang, Dongyuan Li, Peide Zhu, Xiaohui Liang, Xiaochuan Wang, and Manabu Okumura. 2024. Towards cross-modal point cloud retrieval for indoor scenes. In *International Conference on Multimedia Modeling*. Springer, 89–102.

[41] Yuan Zhou, Qiuyue Wang, Yuxuan Cai, and Huan Yang. 2024. Allegro: open the black box of commercial-level video generation model. (2024). https://arxiv.org/abs/2410.15458 arXiv: 2410.15458 [cs.CV].