AgriMus: Developing Museums in the Metaverse for Agricultural Education

Ali Abdari[1,2,*], Alex Falcon[1] and Giuseppe Serra[1]

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

Abstract

Learning agricultural practices—such as gardening, maintaining fruit trees, and general farming techniques—has increasingly shifted towards digital platforms, with tutorials on YouTube being a popular resource. As the metaverse expands, immersive experiences are emerging as powerful tools for skill acquisition. This work introduces AgriMus, a search tool designed for metaverse environments, enabling users to discover both videos and interactive experiences tailored to teaching practical skills in agriculture. AgriMus aims to connect users with relevant virtual spaces where they can learn and practice agricultural tasks in a hands-on, engaging way. Initial experiments conducted on 83 exhibitions demonstrate the potential of zero-shot search methods, achieving 27% R@1, 41% MRR, and 52% nDCG@5. The results also highlight the importance of leveraging the hierarchical structure of exhibition data and integrating state-of-the-art vision-language models to improve search performance. The source code and data of this work is available at https://github.com/aliabdari/AgriMus.

Keywords

Metaverse, Digital Museums, Agriculture Education, Cross-modal Retrieval, Multimedia

1. Introduction

Nowadays, with the user-generated content uploaded on the Internet increasing dramatically every year, it is becoming a common practice to acquire new skills by watching tutorials on video sharing platforms such as YouTube. These tutorial videos span a broad range of different skills, including general life skills such as cooking, home organization, and DIY crafts; technical skills like coding, graphic design, and video editing; and practical hands-on activities like gardening, farming, and maintaining fruit trees. For instance, users can find step-by-step guides on planting and cultivating vegetables, pruning fruit trees for optimal growth, designing irrigation systems, and even employing modern farming technologies, e.g., hydroponics or drone-assisted crop monitoring. This vast repository of user-generated content empowers individuals to learn both everyday and specialized skills at their own pace.

With the rapid growth of the metaverse, a new dimension of learning and skill acquisition is emerging, particularly in areas like agriculture. Initiatives such as the Agriscience Metaverse Academy are already leveraging virtual reality (VR) to provide immersive educational experiences for agriculture teachers and students, enabling them to explore agriscience concepts without the constraints of physical resources. Similarly, projects like "Georgia Agriculture in the Metaverse" introduce AI-powered, game-based learning environments where users can grow crops, manage agricultural businesses, and gain practical farming skills through interactive simulations. These examples illustrate how the metaverse is transforming traditional tutorial-based learning into dynamic, hands-on experiences, making skill development more accessible, engaging, and impactful.

To take advantage of the strengths of both traditional tutorial videos and immersive metaverse experiences, we introduce the AgriMus project, the overview of which can be seen in Figure 1. AgriMus focuses on developing a specialized search tool that empowers users interested in learning agricultural activities to explore and identify relevant agricultural metaverses. By integrating video content with

IRCDL 2025: 21st Conference on Information and Research Science Connecting to Digital and Library Science, February 20–21 2025, Udine, Italy

*Corresponding author.
abdari.ali@spes.uniud.it (A. Abdari); falcon.alex@spes.uniud.it (A. Falcon); giuseppe.serra@uniud.it (G. Serra)
0000-0002-4482-0479 (A. Abdari); 0000-0002-6325-9066 (A. Falcon); 0000-0002-4269-4501 (G. Serra)

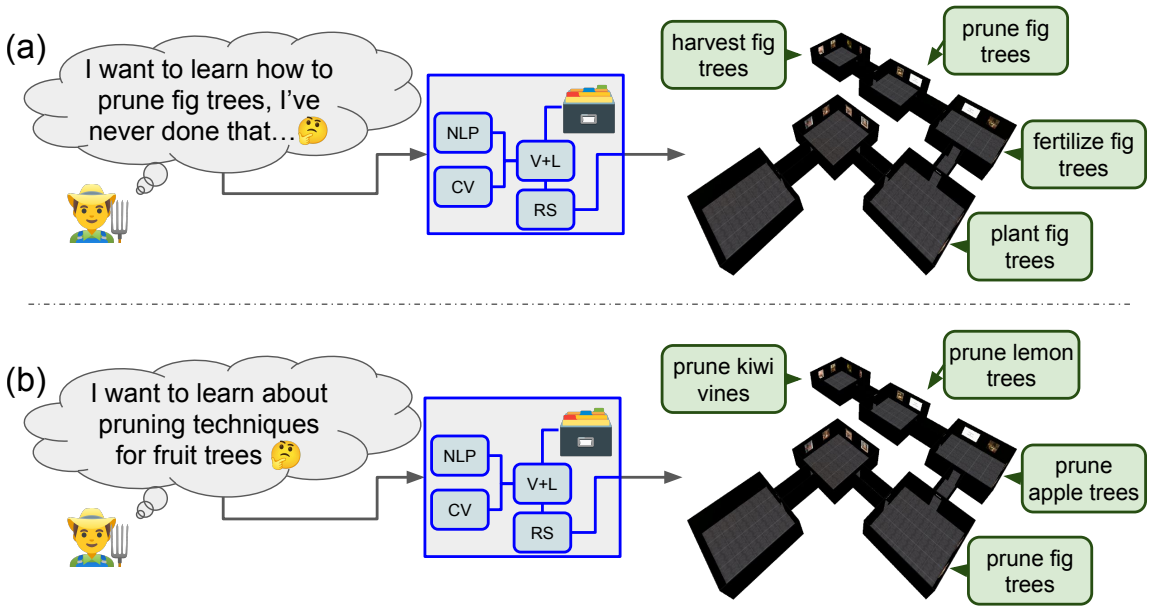

Figure 1: Given the user query, formulated in natural language, the method uses natural language processing (NLP) for it, then combines computer vision (CV) techniques and multimodal analysis (V+L) to process the metaverses available in the database. Then, it recommends (RS) a ranking list of the relevant metaverses. The two cases show possible results. (a) Metaverse focusing on the specific tree (fig), with the rooms dedicated to different aspects for it. (b) Metaverse focusing on the action (pruning), with the rooms dedicated to applying it in diverse agriculture scenarios.

interactive virtual experiences, this tool allows users to search for and access metaverse environments tailored to their specific interests, such as gardening, farming techniques, or advanced agricultural practices. AgriMus bridges the gap between conventional online tutorials and the growing potential of the metaverse, offering a comprehensive platform for skill development in agriculture.

To demonstrate the feasibility of AgriMus, we collected a dataset specifically designed for proof-of-concept purposes. The dataset comprises 83 topical exhibitions, each dedicated to a broad agricultural theme (e.g., pruning fruit trees), with individual rooms focusing on more specific subtopics (e.g., pruning lemon trees). We conducted experiments in a zero-shot scenario, leveraging the hierarchical structure of the exhibitions to model the data as envisioned for AgriMus. Our experimental results demonstrate promising performance, achieving 27% recall at rank 1 (R@1), 66% recall at rank 5 (R@5), and a mean reciprocal rank (MRR) of 41%. Additionally, we achieved 52% normalized discounted cumulative gain (nDCG) at rank 5 and 56% recall at rank 10. These results highlight the effectiveness of the hierarchical approach and validate the potential of AgriMus for enabling efficient exploration and retrieval in agricultural metaverses.

2. Related work

2.1. Digital museums

The emergence of digital museums represents a transformative shift in how cultural heritage is accessed and experienced, offering unprecedented opportunities for engagement and education [1]. With the advancements in technologies such as high-quality 3D modeling and virtual reality (VR), digital museums are becoming more popular and it is possible for them to host rich and immersive experiences. For instance, they allow for detailed representations of artifacts and exhibitions [2, 3], enabling visitors to explore diverse themes ranging from ancient civilizations to contemporary art [4, 5]. Moreover, unlike traditional museums, which are constrained by physical space and operating hours, digital museums can operate continuously, providing access to global audiences at any time.

Thus, digital museums play a vital role in preserving and promoting cultural heritage by making artifacts and traditions accessible to wider audiences. However, they usually focus their attention to cultural heritage. Conversely, this work builds on the concept of digital museums by focusing on the integration of agricultural knowledge and training materials into museum-like exhibits, creating a unique training avenue for novices and practitioners in agricultural domains, which has not been studied in the researches so far. The aim is to support the acquirement of new skills by mixing lecture-like videos and virtual hands-on practice by means of VR experiences.

2.2. Multimedia-rich 3D scenarios

Recent advancements in vision and language techniques have significantly enhanced the retrieval of 3D scenes and objects through natural language queries. The integration of dense captioning methods with RGB-D scans enables the generation of detailed, context-aware descriptions of localized objects within 3D environments [6]. These approaches allow users to input natural language queries to retrieve specific objects or scenes, thereby improving the efficiency and accuracy of retrieval systems. By combining language and 3D visual data, these techniques facilitate more intuitive interactions between humans and machines, enabling natural language descriptions to guide the search and discovery of relevant 3D models or environments.

Instead of focusing on single objects, recent research has focused on more complex indoor scene retrieval using text, involving longer descriptions, as they need to describe many objects and their position within the entire scene. Several contributions were made in this direction, including CRISP [7], which provides a large collection of 3D indoor scenes and their corresponding textual descriptions, Farmare [8] and Adoctera [9] which focus on learning to search furnished multi-room apartments and rank them against user queries. More recently, Text2SceneGraphMatcher [10] introduced a method for aligning open-set text queries with 3D scene graphs to facilitate effective scene retrieval.

However, the approaches mentioned above do not consider the possibility of having inside the scenes some multimedia content which affects the relevance to the user query. This problem raises additional challenges as both the global structure and the local components need to be accounted for in the learned representation in order to fully capture the contents of the scenes and align them to the queries. For instance, in our previous works we investigated the use of cross-modal approaches to rank 3D scenarios comprising additional multimedia data in the form of either videos [11] or images [12].

3. AgriMus: An overview of the project

This section offers an overview of the plans to implement the AgriMus project. These are also presented graphically in Figure 2. The project will involve three main steps, namely data collection, data modeling, and the evaluation phase with an emphasis on user studies.

3.1. Collecting the data

The data collection phase will involve three main ingredients: tutorial videos, experiences, and related descriptions.

For videos, we will use an automated pipeline to collect relevant tutorial videos from YouTube by querying for keywords related to agricultural skills, gardening, and DIY projects. Videos with informative titles will be prioritized to ensure the relevance of the content. The audio tracks of these videos will be transcribed using Whisper [13], a state-of-the-art speech-to-text model known for its high accuracy across multiple languages and challenging audio conditions. The resulting transcripts will serve as a basis for generating detailed textual descriptions. We will use large language models (LLMs) to process these transcripts, as previously done in recent research [14, 15], and extract key procedural steps to produce structured descriptions that enhance video indexing and facilitate the search process.

The process of gathering virtual experiences will involve a combination of automated and manual curation. We will systematically review academic literature to identify virtual agricultural training

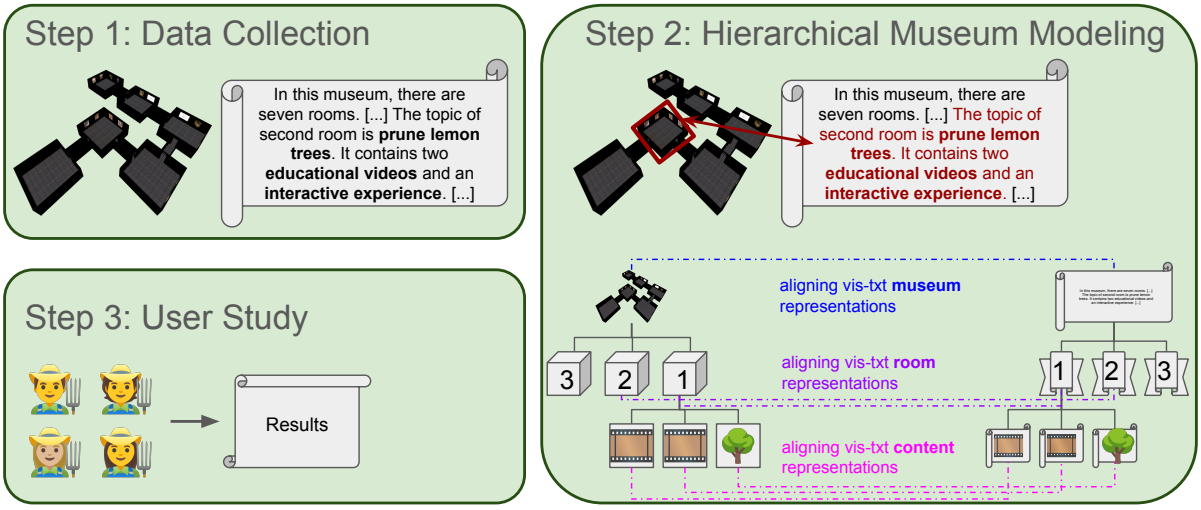

Figure 2: An overview of the AgriMus project. It consists of three main steps. Step 1 is about collecting the required data, comprising topical 3D exhibitions adorned with educational videos and experiences in fields related to agriculture. Step 2 introduces a hierarchical methodology for aligning the visual contents to the textual ones, and also for modeling the exhibitions, with the aim of garnering information about the single experiences or videos, how these form the contents of a room with a specific topic (e.g. how to prune a specific type of tree), and finally how the rooms capture a more comprehensive view on it (e.g. pruning that type of tree, and also growing, harvesting, etc). Step 3 will involve user studies to better understand the user needs and the effectiveness of the proposed methodology in capturing them.

environments described in research papers, with particular attention to interactive simulations and metaverse-based experiences. For instance, Fabrika et al [16] developed a system for educating the user into thinning practices, fundamental for forestry management, whereas even better digital twins of forests were recently created using data and procedural approaches, e.g. [17, 18]. Another example is related to teaching the users to detect ripe fruit, e.g. strawberries [19]. In addition, publicly available amateur simulations and virtual environments created by independent developers will be sourced from online repositories and virtual experience platforms. This dual approach ensures a diverse collection of virtual experiences, covering both high-fidelity simulations and more accessible, grassroots solutions. The collected experiences will be cataloged and integrated into the AgriMus platform, enriching the learning ecosystem with practical, hands-on tools.

3.2. Modeling the museums

The exhibitions collected in the previous step are quite rich in content: each exhibition contains multiple rooms, each containing different videos or experiences. To encode all this information in a way that it is easily searchable and avoids information loss, we will rely on a combination of state-of-the-art computer vision, natural language processing, and multimedia analysis techniques. Specifically, as shown in Figure 2, we plan to use hierarchical modeling to leverage the structure of the exhibitions, roughly divided into content-level (videos or interactive experiences), room-level, and museum-level. By aligning the visual and textual representations within each level (i.e., a video/experience with its description, a room with the descriptions of all its contents, and finally the museum with the full description), it will become easier for the model to learn how to orderly encode them while minimizing information loss [20, 21, 22].

For content-level representations, given that both videos and more complex interactive experiences will be integrated, a mixture of spatial and spatio-temporal models will be used. This will include 2D Large Vision-Language Models (LVLM) such as CLIP [23] and Mobile-CLIP [24], and spatio-temporal LVLMs such as LaViLa [25] or InternVideo2 [26]. In this way, it will be possible to separately encode both

appearance and motion information, useful to better understand the primary entities of the experiences (e.g. the tree species) and the actions performed on it.

For room-level representation, a naive solution would be to aggregate the content representation through mean pooling, eventually learning the weight of each. Alternatively, graph networks could also play an important role in understanding how to aggregate them by capturing relationships and dependencies between contents, at the cost of more computational resources. These have been previously used to capture single objects inside rooms (e.g. furniture) and their relationships by using scene graphs [27, 28].

Finally, for the museum-level representation, different types of aggregation could be used depending on the constraints to be imposed on the exhibition itself. Generally, learning a weighted mean of the room representations could suffice, as the information coming from each room would have its weight defined on the content without, for instance, any constraint on the visit order. However, it is common for exhibitions to have a predefined visit order, usually done by the exhibition curator. Therefore, exploring sequential models (e.g., standard recurrent networks such as LSTM and GRU, or the more recent xLSTM [29] and minGRU [30]) for the aggregation of the rooms could play an important role in how to encode their content into the museum representation. As in the previous case, graph neural networks could also be used to capture neighbor relations between rooms and assess the relevance of each.

3.3. Searching through the museums

Once the representations for the museums are computed, they can be searched using similarity-based approaches. Here, two methodologies can be followed.

As content-level representations involve LVLMs, processing the user queries through the same techniques means that the query representation falls into the same latent space, hence enabling training-free search. However, this would imply either that the museums are modeled without relying on the hierarchy or that the aggregation functions are not trained (e.g., mean or max pooling). Although both cases are likely leading to poorer performance compared to a solution using trained components, they enable effective solutions even in scarce data scenarios. In Section 4, we show some early results obtained using this methodology.

In general, user queries may also be long and articulated, describing specific scenarios and thus requiring more advanced query processing. While large vision-language models (LVLMs) are typically trained with simple captions—often composed of primary entities and a few additional descriptive words (e.g., half of the captions in LAION-2B are less than 50 characters long [31])—there are LVLMs trained to handle more complex query scenarios. An example is represented by LaCLIP [32], which uses Large Language Models to rewrite the original captions paired with the training images. This suggests that the zero-shot approach should also work for longer queries, although it is generally unlikely to perform similarly to a model trained specifically for the task at hand. In particular, training the proposed method using the vision and language data collected in the previous step allows the models to become more tailored to the task, potentially preserving more details in the encoding.

4. Early experimental results

As a proof-of-concept for the AgriMus project, we collected a dataset of exhibitions for educational purposes in the agriculture domain. The details of the dataset are provided in Section 4.1, whereas early experimental results are reported in Section 4.2.

4.1. Collected data

As mentioned above, a staple of the AgriMus project will be the availability of museums, or exhibitions, about important topics for education in the field of agriculture, which we will collect because this is not currently available. For an early prototype of the proposed AgriMus project, we created a set of 83

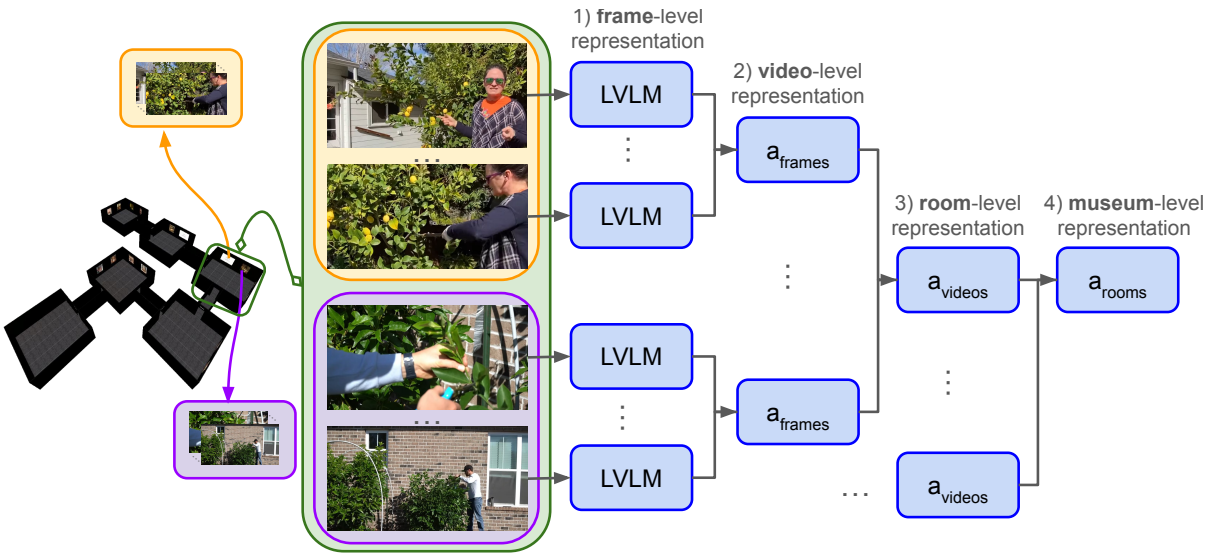

Figure 3: An overview of the prototype for zero-shot understanding of the exhibition contents used in this paper. Starting from the full museum, it highlights one of the rooms (in green) and two of the videos contained in it (yellow and purple). 1) The frames of the videos are processed using a Large Vision-Language Model (LVLM). 2) The frames' representations are then aggregated using the function a_{frames}. 3) The videos are then aggregated using a_{videos} to capture the contents of the room. 4) Finally, a_{rooms} aggregates the rooms' contents to capture the full exhibition. This final representation is then used to rank the exhibitions against the representation of the user query.

topical museums, each focusing on a branch of topics relevant to agricultural education, e.g. tutorials on pruning trees. Then, each room focuses on more specific topics, e.g. how to prune lemon trees. On average, there are 4.6 rooms per museum, with about 11.2 videos per museum. To achieve this, we first collected a total of 288 relevant videos from the HowTo100M dataset [33]. The main topics distilled from the videos range from teaching the user the best practices for growing a tomato plant at home to watering indoor plants or pruning outdoor trees. The topics are extracted using KeyBert [34] looking for representative bigrams in the video title. Examples of topics include keywords for actions such as "sow" and "prune", for entities such as "rose" and "garden", and also for some technical approaches such as "hydroponic". As we looked for bigrams, these are typically grouped in pairs, e.g. "rid" with "weed". In total, we extracted 213 topics. Most of them (about 80) are only bound to one museum or two museums (about 100), and only seven are repeated in four or five museums (Figure 4). The videos, with a length spanning from 38 seconds to 31 minutes, are then "grouped" to form viable candidates' pools for the museum rooms. Specifically, we first selected part of the bigrams (e.g. "growing") to decide a topic for the museum, and then built the rooms based on the second part (e.g. "tomatoes" for one room, and "potatoes" for another room).

4.2. Zero-shot search method

As the dataset collected is small, experiments that involve training the neural network outlined in Section 3 would be unfeasible. Therefore, we designed a zero-shot methodology based on the discussion in Section 3.3. An overview of the zero-shot methodology is illustrated in Figure 3. It is made of four main steps.

First, in each video within the room, 150 frames are uniformly sampled and resized to (H, W), then processed through a spatial LVLM. In the experiments in the following sections, three LVLMs are considered: CLIP [23], Mobile-CLIP [24], and BLIP [35]. H and W are set to 224 for CLIP and BLIP, whereas 256 is used for Mobile-CLIP.

The frame representations are then aggregated by a_{frames}, implemented in the experiments as mean, maximum, or median pooling. Although the mean pooling of frame vectors is quite typical to obtain

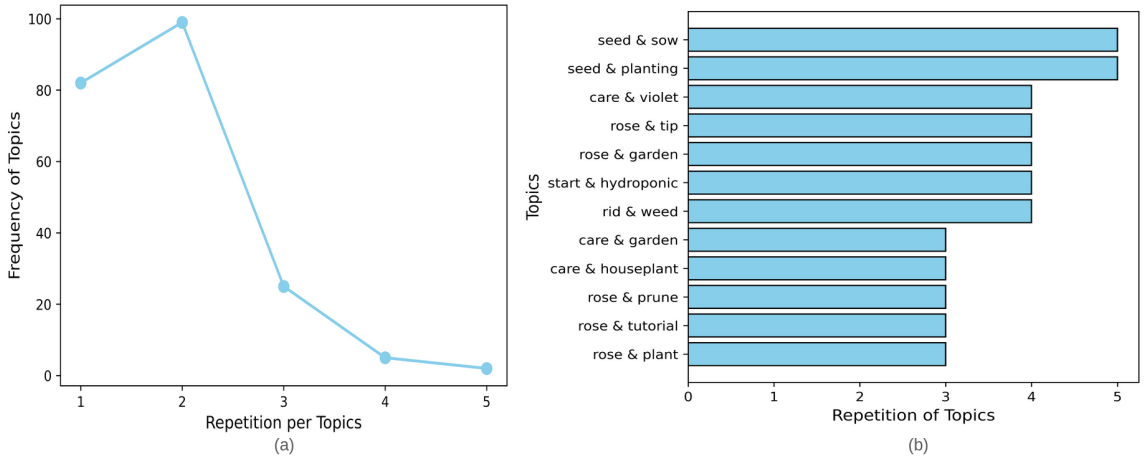

Figure 4: Statistics of the dataset collected. (a) shows the repetition per topics, illustrating that most of the topics have been used in one or two museums, while a few topics have been used in four or five museums. (b) shows some of the topics which have been presented in three or more museums.

a rough representation of the video [36, 37], maximum pooling is another way to aggregate frames by looking at spikes in the features (e.g., often done when reducing the spatial dimensions in deep convolutional networks such as ResNet). However, to avoid overemphasizing spurious spikes which can happen with max pooling, and to avoid diluting meaningful features with mean pooling, which happens especially when the videos are long, median pooling can be a viable candidate as it focuses on the middle value in a region, improving its robustness to extreme values [38].

For the room-level representation, the function a_{videos} is used. As in the previous case, mean, maximum, and median pooling can be used to implement such a function. Although it can be argued that mean and median are more reasonable, as the videos in the room follow the same topic, there are nuances which could be more important to retain. This is the case of many tutorial videos which are longer than the average because they explain how to perform more than one task at once, for instance, showing both how to plow, sow, and water a crop. Therefore, maximum pooling is also a viable candidate for a_{videos}.

Finally, for museum-level aggregation we rely on mean pooling to implement a_{rooms}, so that each room has the same weight in the final encoded representation.

Since we leverage LVLMs to process the visual information, the queries are also processed and encoded through the same models without any additional training. This is because their embedding space is learned by jointly training the visual encoder and aligning it to the textual encoder, so that both output a similar representation for aligned inputs (e.g., an image and its textual description). In our setting, the test queries are made of bigrams which consist of the 213 topics extracted from the video titles. To perform the search, the queries are first tokenized and encoded through the textual encoder of the LVLM, and then cosine similarity is used to rank the museum representations created by a_{rooms}.

4.3. Evaluation metrics

To assess the performance of the system, a relevance score was computed for each exhibition given a query q. The score for museum m is a real value computed by summing 1.0 for each room in m that has q as one of its topics, and 0.1 for each video in other rooms which has q as one of its topics. For instance, if the query is “rid weed” and a museum has two rooms, one with topics “rid weed” and “start hydroponic”, and the other one with “rid rose”. In the second room there are four videos inside, two of which have “rid weed” in their topics (note that one video may have more topics extracted from it). Then, the relevance score of m to q is 1.2 as 1.0 is summed for the first room, and 0.2 is summed for the two relevant videos in the second room. When computing the recall rates and the median rank, the relevant museums are those for which the relevance score is the highest in the ranking list, for that

Table 1
We investigate different aggregation styles for the functions a_{frames}, a_{videos}, and a_{rooms}. CLIP is used as LVLM to process and encode the video frames. Discussion in Section 4.4.

Aggregation of			Recall						
Frames	Videos	Rooms	R@1	R@5	R@10	MedR	MRR	nDCG@5	nDCG@10
Mean	Mean	Mean	**23.94**	53.05	**70.89**	**5**	**39.09**	51.77	55.34
Median	Mean	Mean	20.18	**53.52**	70.42	**5**	36.85	**52.74**	**55.96**
Max	Mean	Mean	7.51	27.23	45.53	12	18.50	44.63	52.83
Mean	Median	Mean	19.71	50.70	69.48	**5**	36.11	50.35	54.09
Median	Median	Mean	19.24	50.70	69.95	**5**	35.09	51.12	54.74
Max	Median	Mean	10.32	26.76	38.02	19	19.18	43.21	49.59
Mean	Max	Mean	20.65	41.31	57.74	7	31.57	48.94	54.28
Median	Max	Mean	18.77	40.84	53.99	9	30.11	49.41	54.37
Max	Max	Mean	11.73	39.98	45.53	12	22.35	49.17	58.15

query.

The performance evaluation is done using four main metrics: Recall at rank K (R@k), Median rank (MedR), Mean Reciprocal Rank (MRR), and Normalized Discounted Cumulative Gain at rank k(nDCG@k). R@k measures the proportion of relevant museums found within the top k retrieved items. MedR represents the median rank position of the first relevant item across queries. MRR evaluates the rank position of the first relevant item, averaging the reciprocal of the rank across queries. nDCG assesses the quality of the ranking list, with higher-ranked relevant items contributing more to the score, rewarding systems that prioritize important results. In all metrics apart from Median rank, the higher value the better performance.

4.4. Which aggregation style is best?

As mentioned, there are several reasons supporting the use of mean, maximum, or median pooling to implement the functions a_{frames}, a_{videos}, and a_{rooms} in the zero-shot search method explored in this paper. Here, we explore several combinations of these functions and assess their performance on the dataset collected. The results are reported in Table 1.

First, aggregating the frames using mean leads to the best R@1 and MRR both when using mean (23.94% R@1 and 39.09% MRR), median (19.71% and 36.11%), or maximum pooling (20.65% and 31.57%) to aggregate the videos, compared to using median or max. In particular, the difference in performance with max pooling is ample compared to median pooling. On the one hand, it shows that preserving some information from all the frames, although noisily, is effective in this scenario. On the other hand, it confirms that using maximum pooling becomes too sensible to spurious spikes and possibly loses sight of the general content of the video, leading to the worst results, e.g. 7.51% R@1 and 18.50% MRR in the case of (max, mean, mean).

Second, using mean pooling for all three functions, i.e. the row represented by (mean, mean, mean), leads to 23.94% R@1 and 39.09% MRR, whereas all the other combinations have less than 20% R@1 and 37% MRR. It also achieves 51.77% nDCG@5 and 55.34% nDCG@10, which ranks second in our experiments as (median, mean, mean) achieves 52.74% nDCG@5 and 55.96% nDCG@10. This indicates a higher chance to retrieve a relevant museum in the first rank than other combinations and a good quality of the proposed ranking lists, hence representing a good candidate for the proposed zero-shot method. Therefore, in the following experiments we used (mean, mean, mean).

4.5. Which feature extractor is best?

In the previous experiment, using mean pooling for all three aggregation functions atop CLIP frame features led to the best results. Here, we explore how other LVLMs affect the final performance of our zero-search method. Specifically, we test Mobile-CLIP [24] and BLIP [35], and combinations of two to three LVLMs by concatenating the frame features. The results are reported in Table 2.

Table 2
We investigate different LVLMs and their combination to extract the frame-level features. The aggregation functions are set to mean pooling. Discussion in Section 4.5.

Feature extractor	R@1	R@5	R@10	MedR	MRR	nDCG@5	nDCG@10
CLIP	23.94	53.05	70.89	5	39.09	51.77	55.34
BLIP	0.46	3.28	12.20	44	4.99	44.38	50.31
Mobile-CLIP	**27.23**	**56.33**	**75.58**	**4**	**41.33**	52.55	56.57
CLIP+BLIP	20.65	47.88	68.07	6	34.76	54.59	**58.85**
CLIP+MCLIP	22.06	53.52	71.83	5	38.44	51.59	55.10
BLIP+MCLIP	7.98	17.84	29.10	26	15.04	48.54	53.59
CLIP+BLIP+MCLIP	20.65	48.82	68.05	6	34.99	**54.60**	58.56

Table 3
We validate the assumption that, even when performing zero-shot search, leveraging the hierarchical nature of the data is useful. Mobile-CLIP is used as the LVLM for frame features extraction, and the aggregation functions are set to mean pooling for our approach. Discussion in Section 4.6.

Feature extractor	R@1	R@5	R@10	MedR	MRR	nDCG@5	nDCG@10
Hierarchical (ours)	**27.23**	**56.33**	**75.58**	**4**	**41.33**	**52.55**	**56.57**
Video-level (mean frames, mean videos)	26.29	55.39	**75.58**	**4**	40.34	52.50	56.48
Video-level (max frames, mean videos)	11.73	29.57	38.49	16	21.20	49.06	55.92
Video-level (max frames, max videos)	8.92	22.06	30.51	21	17.08	43.80	49.35

First, using Mobile-CLIP led to an increase in performance compared to CLIP, for instance from 23.94% R@1 and 39.09% MRR to 27.23% and 41.33%.

Second, combining the information extracted by the LVLMs does not lead to better results. Specifically, with two methods the best results are obtained by CLIP+Mobile-CLIP, but they still fall short of Mobile-CLIP on its own, for instance their combinations obtains 22.06% R@1 and 38.44% MRR, yet these are lower than those obtained by Mobile-CLIP (27.23% and 41.33%). Although putting together all the models leads to slightly better nDCG compared to Mobile-CLIP (e.g. 54.60% nDCG@5 compared to 52.55%), the increased computational or storage costs would not make the solution better.

4.6. Is an hierarchical approach better than a flat one?

For the future of the AgriMus project, we hypothesized that leveraging the hierarchical nature of museums is fundamental to correctly model them, both when training the components and when performing zero-shot search. Here, we validate such hypothesis by performing the aggregation of all the videos in the museum at the video level, neglecting the room separation. The LVLM is set to Mobile-CLIP and the aggregation functions to mean pooling, as this combination performed best in the previous experiments. The results are reported in Table 3.

The main result is a confirmation of the hypothesis, as leveraging the hierarchy leads to 27.23% R@1, 41.33% MRR, 52.55% nDCG@5, and 56.57% nDCG@10, whereas in the other ablations, the best results are 26.29% R@1, 40.34% MRR, 52.50% nDCG@5, and 56.48% nDCG@10. Although the use of maximum pooling leads to significantly worse results, the use of mean pooling at the video level leads to comparable results to the proposed method under several metrics, especially those looking above the first rank (R@5 and 10, nDCG@5 and 10). Nonetheless, we hypothesize that training the aggregation functions will lead to considerably better performance, as that would allow better preservation of the temporal information in the videos and improve the encoding capabilities for the videos and the rooms.

5. Discussion/limitations/future work

In this section, we highlight the limitations of our current approach and outline directions for future work.

As the current implementation of AgriMus relies on a zero-shot search method, we employed simple aggregation operations to combine the representations of frames, videos, and rooms. While this approach is straightforward and computationally efficient, it is well known that such operations are not optimal, and for instance they tend to lose temporal information in videos [39, 40]. In future iterations, once we have collected a sufficient amount of data, we plan to experiment with neural sequential models and learned aggregation functions. These should enhance the system's ability to recognize temporal patterns, leading to better video representation and improved search accuracy. Training on larger datasets will not only improve content recognition but also facilitate a deeper usage of the hierarchical structure present in the exhibitions, contributing to more precise search results.

Another challenge is the inherent diversity and complexity of topics related to agriculture, gardening, and related fields. These domains encompass a wide range of subfields, each requiring specific expertise and datasets. To develop a robust and comprehensive system useful to both practitioners and novices, it is essential to collect a larger and more diverse set of videos. For example, there are currently no videos covering certain tree species, such as cedar trees. Interestingly, increasing the scope of the dataset could also facilitate the creation of more specialized virtual museums. For instance, an exhibition might focus specifically on "lemon trees", with rooms dedicated to different stages of growth and care (e.g., planting, watering, pruning, harvesting). Alternatively, broader topics like "growing vegetables indoors" could be broken down into rooms focusing on various crops, such as tomatoes, potatoes, and zucchini. This structured, hierarchical approach will enhance the learning experience by organizing content logically and progressively.

In addition to expanding the video dataset, future efforts will focus on incorporating virtual experiences that allow users to practice within the metaverse. By complementing tutorial videos with interactive, immersive environments, users can engage more deeply with the content, reinforcing their learning through hands-on experiences. Such experiences will be particularly valuable for tasks that require manual skills, such as pruning or grafting, as they enable users to practice techniques in a simulated environment. User studies need also to be conducted to assess the comprehensiveness of the exhibitions and their educational effectiveness.

6. Conclusions

With the growth of the internet and user-generated content, video tutorials have become essential tools for supporting educational efforts across various domains, teaching the watchers best practices to grow vegetables at home, prune fruit trees, and other practical agricultural skills. As the metaverse continues to evolve, these video tutorials can be complemented by interactive and immersive experiences, enhancing the learning process by providing hands-on practice opportunities.

To realize this vision, we introduced the AgriMus project, which focuses on developing digital exhibitions aimed at educating both novices and practitioners in a broad range of topics related to agriculture and gardening. AgriMus aims to build a search tool that allows users to explore these virtual museums, enabling them to watch tutorial videos to learn best practices and then engage in interactive experiences to practice and consolidate their skills within the metaverse.

As an initial step, we collected a dataset of 83 exhibitions, each consisting of multiple topical rooms enriched with video content. We conducted zero-shot experiments, achieving 27.23% R@1, 75.58% R@10, 41.33% MRR, and 52.55% nDCG@5 on a test set of 213 queries. Our experimental results demonstrated that leveraging the hierarchical structure of the data improves performance. In addition, they validated design choices for our scenario: mean pooling proved to be the most effective aggregation method, and Mobile-CLIP outperformed other models in feature extraction from video frames.

Looking ahead, several steps remain to fully realize the AgriMus project. We plan to expand the dataset by incorporating more videos to capture greater diversity across agricultural topics. Furthermore, integrating temporal information will enhance video content representation, improving search accuracy and museum organization. Lastly, conducting user evaluations will be crucial to refining the system and ensuring its effectiveness in real-world scenarios.

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] 3D-Ace, What is a virtual museum: Benefits, types and creation process, 2022. URL: https://3d-ace.com/blog/virtual-museum/, accessed: 2024-12-23.

[2] C. Kiourt, A. Koutsoudis, G. Pavlidis, Dynamus: A fully dynamic 3d virtual museum framework, Journal of Cultural Heritage 22 (2016) 984–991.

[3] E. Zidianakis, N. Partarakis, S. Ntoa, A. Dimopoulos, S. Kopidaki, A. Ntagianta, E. Ntafotis, A. Xhako, Z. Pervolarakis, E. Kontaki, et al., The invisible museum: A user-centric platform for creating virtual 3d exhibitions with vr support, Electronics 10 (2021) 363.

[4] M. Barszcz, K. Dziedzic, M. Skublewska-Paszkowska, P. Powroznik, 3d scanning digital models for virtual museums, Computer Animation and Virtual Worlds 34 (2023) e2154.

[5] M. Merella, S. Farina, P. Scaglia, G. Caneve, G. Bernardini, A. Pieri, A. Collareta, G. Bianucci, 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 (2023) 6762–6776.

[6] Z. Chen, A. Gholami, M. Nießner, A. X. Chang, Scan2cap: Context-aware dense captioning in rgb-d scans, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, 2021, pp. 3193–3203.

[7] F. Yu, Z. Wang, D. Li, P. Zhu, X. Liang, X. Wang, M. Okumura, Towards cross-modal point cloud retrieval for indoor scenes, in: International Conference on Multimedia Modeling, Springer, 2024, pp. 89–102.

[8] A. Abdari, A. Falcon, G. Serra, 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, 2023, pp. 4293–4303.

[9] A. Abdari, A. Falcon, G. Serra, Adoctera: Adaptive optimization constraints for improved text-guided retrieval of apartments, in: Proceedings of the 2024 International Conference on Multimedia Retrieval, 2024, pp. 1043–1050.

[10] J. Chen, D. Barath, I. Armeni, M. Pollefeys, H. Blum, “where am i?” scene retrieval with language, in: European Conference on Computer Vision, Springer, 2025, pp. 201–220.

[11] A. Abdari, A. Falcon, G. Serra, Metaverse retrieval: Finding the best metaverse environment via language, in: Proceedings of the 1st International Workshop on Deep Multimodal Learning for Information Retrieval, 2023, pp. 1–9.

[12] A. Abdari, A. Falcon, G. Serra, A language-based solution to enable metaverse retrieval, in: International Conference on Multimedia Modeling, Springer, 2024, pp. 477–488.

[13] A. Radford, J. W. Kim, T. Xu, G. Brockman, C. McLeavey, I. Sutskever, Robust speech recognition via large-scale weak supervision, in: International conference on machine learning, PMLR, 2023, pp. 28492–28518.

[14] W. R. Huang, C. Allauzen, T. Chen, K. Gupta, K. Hu, J. Qin, Y. Zhang, Y. Wang, S.-Y. Chang, T. N. Sainath, Multilingual and fully non-autoregressive asr with large language model fusion: A comprehensive study, in: ICASSP 2024-2024 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP), IEEE, 2024, pp. 13306–13310.

[15] R. Ma, M. Qian, P. Manakul, M. Gales, K. Knill, Can generative large language models perform asr error correction?, arXiv preprint arXiv:2307.04172 (2023).

[16] M. Fabrika, P. Valent, L. Scheer, Thinning trainer based on forest-growth model, virtual reality and computer-aided virtual environment, Environmental modelling & software 100 (2018) 11–23.

[17] A. S. Badr, D. D. Hsiao, S. Rundel, R. de Amicis, Leveraging data-driven and procedural methods

for generating high-fidelity visualizations of real forests, Environmental Modelling & Software 172 (2024) 105899.
[18] H. Qiu, H. Zhang, K. Lei, H. Zhang, X. Hu, Forest digital twin: A new tool for forest management practices based on spatio-temporal data, 3d simulation engine, and intelligent interactive environment, Computers and Electronics in Agriculture 215 (2023) 108416.
[19] J. J. Chai, J.-L. Xu, C. O'Sullivan, Real-time detection of strawberry ripeness using augmented reality and deep learning, Sensors 23 (2023) 7639.
[20] K. Ashutosh, R. Girdhar, L. Torresani, K. Grauman, Hiervl: Learning hierarchical video-language embeddings, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 23066–23078.
[21] X. Wang, S. Li, K. Kallidromitis, Y. Kato, K. Kozuka, T. Darrell, Hierarchical open-vocabulary universal image segmentation, Advances in Neural Information Processing Systems 36 (2024).
[22] Q. Ye, G. Xu, M. Yan, H. Xu, Q. Qian, J. Zhang, F. Huang, Hitea: Hierarchical temporal-aware video-language pre-training, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, 2023, pp. 15405–15416.
[23] A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, et al., Learning transferable visual models from natural language supervision, in: International conference on machine learning, PMLR, 2021, pp. 8748–8763.
[24] P. K. A. Vasu, H. Pouransari, F. Faghri, R. Vemulapalli, O. Tuzel, Mobileclip: Fast image-text models through multi-modal reinforced training, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2024, pp. 15963–15974.
[25] Y. Zhao, I. Misra, P. Krähenbühl, R. Girdhar, Learning video representations from large language models, in: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2023, pp. 6586–6597.
[26] Y. Wang, K. Li, X. Li, J. Yu, Y. He, G. Chen, B. Pei, R. Zheng, Z. Wang, Y. Shi, et al., Internvideo2: Scaling foundation models for multimodal video understanding, in: European Conference on Computer Vision, Springer, 2025, pp. 396–416.
[27] L. Gao, J.-M. Sun, K. Mo, Y.-K. Lai, L. J. Guibas, J. Yang, Scenehgn: Hierarchical graph networks for 3d indoor scene generation with fine-grained geometry, IEEE Transactions on Pattern Analysis and Machine Intelligence 45 (2023) 8902–8919.
[28] J. Wald, N. Navab, F. Tombari, Learning 3d semantic scene graphs with instance embeddings, International Journal of Computer Vision 130 (2022) 630–651.
[29] M. Beck, K. Pöppel, M. Spanring, A. Auer, O. Prudnikova, M. Kopp, G. Klambauer, J. Brandstetter, S. Hochreiter, xlstm: Extended long short-term memory, Advances in Neural Information Processing Systems (2024).
[30] L. Feng, F. Tung, M. O. Ahmed, Y. Bengio, H. Hajimirsadegh, Were rnns all we needed?, arXiv preprint arXiv:2410.01201 (2024).
[31] C. Schuhmann, R. Beaumont, R. Vencu, C. Gordon, R. Wightman, M. Cherti, T. Coombes, A. Katta, C. Mullis, M. Wortsman, et al., Laion-5b: An open large-scale dataset for training next generation image-text models, Advances in Neural Information Processing Systems 35 (2022) 25278–25294.
[32] L. Fan, D. Krishnan, P. Isola, D. Katabi, Y. Tian, Improving clip training with language rewrites, Advances in Neural Information Processing Systems 36 (2024).
[33] A. Miech, D. Zhukov, J.-B. Alayrac, M. Tapaswi, I. Laptev, J. Sivic, Howto100m: Learning a text-video embedding by watching hundred million narrated video clips, in: Proceedings of the IEEE/CVF international conference on computer vision, 2019, pp. 2630–2640.
[34] M. Grootendorst, Keybert: Minimal keyword extraction with bert., 2020. URL: https://doi.org/10.5281/zenodo.4461265. doi:10.5281/zenodo.4461265.
[35] J. Li, D. Li, C. Xiong, S. Hoi, Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, in: International conference on machine learning, PMLR, 2022, pp. 12888–12900.
[36] M. Bain, A. Nagrani, G. Varol, A. Zisserman, Frozen in time: A joint video and image encoder for end-to-end retrieval, in: Proceedings of the IEEE/CVF international conference on computer

vision, 2021, pp. 1728–1738.
[37] V. Gabeur, C. Sun, K. Alahari, C. Schmid, Multi-modal transformer for video retrieval, in: Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part IV 16, Springer, 2020, pp. 214–229.
[38] W. Shi, C. C. Loy, X. Tang, Deep specialized network for illuminant estimation, in: Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV 14, Springer, 2016, pp. 371–387.
[39] X. Jiang, Y. Gong, X. Guo, Q. Yang, F. Huang, W.-S. Zheng, F. Zheng, X. Sun, Rethinking temporal fusion for video-based person re-identification on semantic and time aspect, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 34, 2020, pp. 11133–11140.
[40] M. Li, H. Xu, J. Wang, W. Li, Y. Sun, Temporal aggregation with clip-level attention for video-based person re-identification, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, 2020, pp. 3376–3384.