A Matter of Time: Revealing the Structure of Time in Vision-Language Models

Nidham Tekaya*
ntekaya@fhstp.ac.at
St. Pölten University of Applied Sciences
St. Pölten, Austria
TU Wien
Vienna, Austria

Manuela Waldner
manuela.waldner@tuwien.ac.at
TU Wien
Vienna, Austria

Matthias Zeppelzauer
mzeppelzauer@fhstp.ac.at
St. Pölten University of Applied Sciences
St. Pölten, Austria

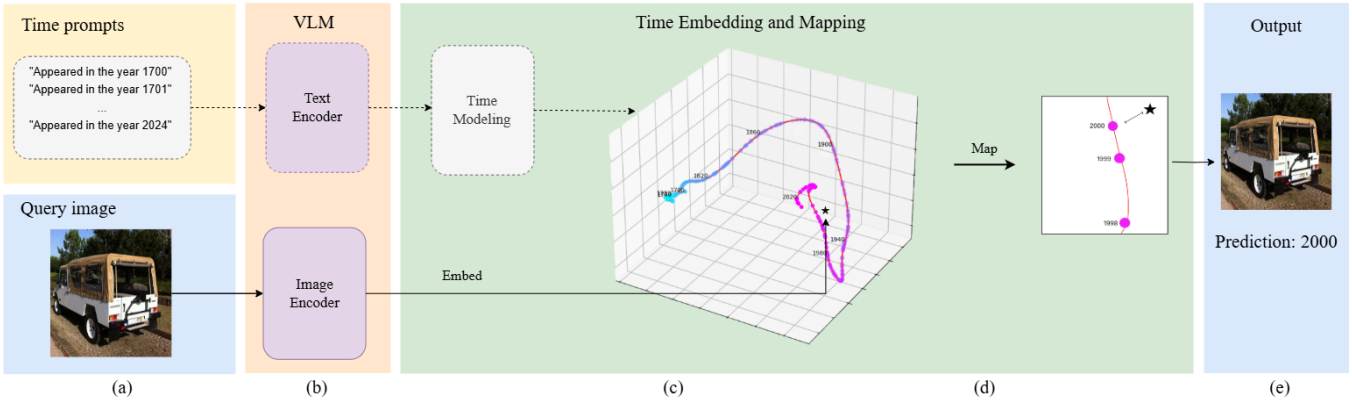

Figure 1: Overview of our approach to time assessment in images. (a) Time-specific prompts and query images are input to the VLM. (b) Text and image encoders generate embeddings in the shared multimodal space. (c) Time embeddings form a chronological manifold that we model as a 1D curve in a subspace of the embedding space. (d) Image embeddings are mapped to the closest year on the time curve. (e) Time predictions (in years) are output for each query image.

Abstract

Large-scale vision-language models (VLMs) such as CLIP have gained popularity for their generalizable and expressive multimodal representations. By leveraging large-scale training data with diverse textual metadata, VLMs acquire open-vocabulary capabilities, solving tasks beyond their training scope. This paper investigates the temporal awareness of VLMs, assessing their ability to position visual content in time. We introduce TIME10k, a benchmark dataset of over 10,000 images with temporal ground truth, and evaluate the time-awareness of 37 VLMs by a novel methodology. Our investigation reveals that temporal information is structured along a low-dimensional, non-linear manifold in the VLM embedding space. Based on this insight, we propose methods to derive an explicit "timeline" representation from the embedding space. These representations model time and its chronological progression and thereby facilitate temporal reasoning tasks. Our timeline approaches achieve competitive to superior accuracy compared to a prompt-based baseline while being computationally efficient. All code and data are available at https://tekayanidham.github.io/timeline-page/.

*Corresponding author.

,
2025.

CCS Concepts

• **Information systems** → **Multimedia and multimodal retrieval**; *Information extraction*; • **Computing methodologies** → **Computer vision representations**.

Keywords

Multimodal representations, Vision-language models, Time modeling, Time reasoning, Time estimation, Benchmark dataset

1 Introduction

Time plays a fundamental role in the analysis of historical image and video collections and represents an essential aspect in cultural studies, historical analysis, and the management of digital archives. Determining when certain artifacts first appeared in time enables answering questions, such as: *how old is this object?* and *when did it first appear?* It further provides valuable constraints for dating historical images, i.e. *what is the earliest possible time this picture was taken?*. From a technical perspective, it is important to examine

whether vision-language models (VLMs) exhibit inherent temporal reasoning. If so, understanding how this reasoning is encoded in their learned representations is key to evaluating their ability to infer the temporal origin of objects.

Despite advances in visual analysis and understanding, dedicated methods for time assessment remain underdeveloped due to time's abstract nature and a lack of datasets with sufficient and adequate metadata. We investigate whether VLMs implicitly capture temporal structures and if they provide a certain awareness to time. To the best of our knowledge, this is the first analysis of this kind.

With the emergence of VLMs being trained on large-scale image datasets [9, 35] and paired with textual metadata, powerful representations have become available. These pre-trained VLMs can map concepts to images and vice versa, facilitating numerous downstream tasks [50]. The metadata used for training VLMs is often derived from web sources, including associated text such as alternative image descriptions from HTML or other contextual annotations [35, 37, 13], which may implicitly encode absolute or relative temporal information like object descriptions (e.g., "a vintage car from the 1960s" or "an early model of a mobile phone"), event contexts (e.g., "photographed during the 1980 Olympic Games"), or references to time periods (e.g., "artwork from the Renaissance period"). While the exact nature and quality of these annotations can vary significantly, we can assume that the learned representations implicitly capture a certain amount of temporal information and context, making the models to some degree time-aware.

This paper is the first to systematically explore the time-awareness and temporal modeling capabilities of VLMs. By analysing their embedding spaces, we investigate how these models implicitly encode temporal information. Our investigation specifically focuses on *human-made artifacts* and on the "*time of first appearance*" defined as the time point when a given object was first introduced in history or became publicly available (e.g., the year when a new car model was released). Note that, our goal is not to achieve perfect accuracy on exact release dates, but more importantly to assess whether models can approximate these temporal landmarks within a reasonable margin, providing meaningful temporal context. In our investigation, we particularly focus on contrastively trained VLMs such as CLIP and its successors [24, 27, 6] with open-vocabulary capabilities. These architectures represent the backbones for most recent foundation VLMs, including also generative models like [39, 21, 26]. To cover a broad spectrum of architectures, we include ImageBind [34], ViT-Lens [42], EVA-CLIP [31, 33], and SigLIP [49].

Our investigation is guided by the following central research questions:

- RQ1: To what extent are open-vocabulary VLMs inherently time-aware?
- RQ2: How is the temporal information of images and text structured in the embedding space?
- RQ3: If temporal structure exists (RQ2), how can we represent time to effectively and efficiently predict the time of first appearance of the images?

There are a number of challenges to tackle to answer our research questions: (i) there is limited availability of datasets with explicit time annotations which makes the systematic comparison and performance benchmarking difficult; (ii) VLM embedding spaces are high-dimensional (typically 512+ dimensions) with latent dimensions that lack semantic interpretation, making it challenging to identify which dimensions encode temporal information; (iii) prompting VLMs for temporal context (i.e. prompt probing [22]) is possible (due to their open vocabulary nature), however, it requires generating embeddings for each year, limiting its scalability for larger temporal ranges.

To address the above challenges, this paper introduces a novel methodology for systematically exploring and leveraging the temporal reasoning capabilities of VLMs. Our main contributions are as follows:

- We introduce **TIME10k**, a temporally annotated dataset with over 10,000 images from 6 classes of objects, enabling systematic evaluation and comparison of VLMs with respect to temporal awareness and time prediction capabilities.
- We propose a comprehensive framework for objectively evaluating time-awareness and investigate 37 state-of-the-art VLMs, examining various backbones, architectures, and prompting strategies, and providing a clear pathway for benchmarking temporal reasoning.
- A novel approach for deriving an explicit model of time from a VLM's embedding space in the form of a sequential "timeline" representation, as illustrated in Figure 1, which can be leveraged to effectively and efficiently place images in a temporal context.

2 Related Work

Related work for our work comes from different areas, briefly reviewed in the following.

Open-Vocabulary Vision-Language Models. VLMs have gained traction since the introduction of CLIP [3], which uses contrastive training to align image-text pairs and enables open-vocabulary learning. The result is a shared latent space where semantically related images and text are closely positioned in terms of dot-product similarity. CLIP has inspired the development of models such as OpenCLIP [29], EVA-CLIP [31, 33], SIGLIP [49], ImageBind [34], and ViT-Lens [42], which integrate textual and visual data and even additional modalities. These models demonstrate strong zero-shot classification [32] and cross-modal retrieval capabilities [8]. Researchers have examined how CLIP embeddings encode perceptual attributes [41] and geometric reasoning [10]. Building upon these lines of investigation, our work analyses whether VLMs implicitly encode temporal information about depicted objects.

Textual Modeling of Temporal Aspects. Recent work has demonstrated that language models learn structured temporal representations from textual data. Gurnee & Tegmark [15] identified "time neurons" that encode temporal coordinates in linear representations when processing text about historical figures, while Heinzerling & Inui [14] found that temporal properties like birth years are encoded in low-dimensional linear subspaces and demonstrated control by editing internal model activations to systematically change the model's temporal outputs. Although highly related to our work, these approaches focus only on the textual modality and LLMs. Our work investigates temporal context across the visual and the textual domain in VLMs. Interestingly, while the

above works find time to be modelled linearly, our work reveals a *non-linear* structure of time in VLMs.

Visual Modeling of Temporal Aspects. Temporal aspects in images have been studied for face aging [54], historical portrait generation [48], and video modeling [47]. Chen et al. [11] used StyleGAN to synthesize faces across historical periods (1880s-2020s) through stylistic transformations. Methods have leveraged explicit temporal metadata from LAION-5B [35, 23, 7, 13] for historical retrieval. Most related is Barancová et al. [4], who explored OpenCLIP for zero-shot dating of historical photographs from 1950-1999 [43], showing bias towards earlier years and improved accuracy via fine-tuned logistic regression. We go beyond this study and investigate *intrinsic* temporal properties across 37 VLMs without fine-tuning and introduce methods for deriving an *explicit* timeline representation from embedding spaces.

Embedding Space Analysis. Understanding high-dimensional embeddings is crucial for the interpretability of the models. To improve it, Concept Activation Vectors (CAVs) [1] define semantic directions in embedding spaces, while PCA-based GANSpace [12] uncovers latent structures related to semantic attributes such as pose, lighting, ageing, and object transformations in generative models. Furthermore, dimensionality reduction techniques such as UMAP [20] have been applied to explore contrastive embeddings, for example in [17] for interactive cluster analysis in contrastive spaces. However, to the best of our knowledge, prior work has not examined whether temporal information manifests in some kind of structure in embedding spaces.

3 Methodology

In the following, we outline our methodology. We first describe *Time Probing*, a baseline for temporal inference (Section 3.1). Next, we analyse the spatial structure of temporal embeddings in the VLM's embedding space to shed light on the time-awareness of VLMs (Section 3.2) and finally we propose a method to derive an explicit representation of time from the embedding space that enables efficient temporal inference as an alternative to time probing (Section 3.3). Note that in the following we refer to *time embeddings* as the outputs of the VLM's text encoder when prompted with time-specific queries, which represent specific years and serve as temporal anchors within the shared latent space.

3.1 Time Probing: A Baseline for Retrieving Temporal Context (RQ1, RQ2)

Time probing is a prompt-based baseline inspired by Barancová et al. [4], who used VLMs for zero-shot dating of photographs. We adapt this approach to estimate when depicted objects first appeared by exploiting the cross-modal alignment in the VLM embedding space (Figure 2). Time probing generates text prompts by combining a template sentence with a year $y \in Y$ for a given time range $Y = [Y_{\text{min}}, \ldots, Y_{\text{max}}]$. To refer to the "time of first appearance", prompts may be defined e.g. as "First appeared in the year y" or "Was introduced in the year y". These prompts are encoded using the VLM's text encoder, resulting in *time embeddings* $\mathbf{T}_y$ for individual years. Similarly, the query image is encoded into an *image embedding* $\mathbf{I}$ using the image encoder. Both embeddings exist in the same shared latent space, enabling direct comparison via

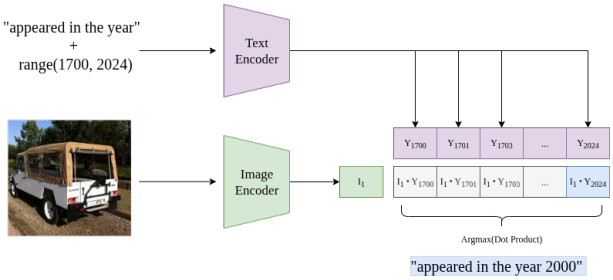

Figure 2: Time probing: temporal inference via dot-product similarity. Time embeddings (here for years 1700-2024) are compared with image embeddings. The time embedding with the highest similarity indicates the year of first appearance.

dot-product similarity. The predicted year y_{pred} is finally derived as the year whose time embedding has the highest similarity:

$$y_{\text{pred}} = \arg\max_{y \in Y} (\mathbf{I}^{\text{T}} \cdot \mathbf{T}_y), \tag{1}$$

where $\mathbf{I} \in \mathbf{R}^N$ is the high-dimensional image embedding, $\mathbf{T}_y \in \mathbf{R}^N$ is the time embedding for year y, and Y is the set of candidate years. Time probing serves as a baseline in our experiments, particularly for answering RQ1. It leverages the cross-modal matching capabilities of VLMs to assess temporal information implicitly captured during training. However, time probing is limited in that it only performs local analysis, treating each year independently without modeling temporal progression or dependencies between time points. It further scales poorly with the number of years in Y.

3.2 Embedding Space Analysis (RQ2)

Time probing treats each year's embedding independently, ignoring potential relationships between time points. This prevents us from capturing higher-level temporal structure. To advance beyond simple probing, we need to understand how time is organised within the VLM's embedding space, which is the central focus of RQ2. The spatial distribution of the time embeddings in the embedding space is a priori unknown. Time points could follow linear or non-linear relationships, form clusters, or lack any coherent structure. To investigate the spatial structure of time embeddings we leverage dimensionality reduction techniques. Thereby, we proceed in two ways: First, we visualize lower-dimensional projections of time embeddings to identify potential structures. Second, we generate one-dimensional projections to measure how well individual years align chronologically, to test whether a temporal order exists among the embeddings. For our analysis, we select the following dimensionality reduction techniques:

Kernel PCA: (KPCA) [2] is chosen for two reasons: First, it enables the use of the cosine kernel, which directly aligns with the metric used in contrastively trained VLMs and thereby ensures consistency with the intrinsic structure of the embedding space. Second, it preserves the global structure of the underlying space and does not introduce local distortions, thereby avoiding overfitting.

UMAP: [20] is a topological dimensionality reduction approach that aims to preserve most of the global structure of the space but

not necessarily the local metric structure. It is thus complementary to KPCA. UMAP actually preserves the topology and tries to discover a manifold on which the input data are uniformly distributed [28]. This makes UMAP a promising approach for testing whether temporal information forms a structured manifold in the embedding space. Additionally, UMAP's ability to transform and embed new data points into a previously computed projection (crucial for projecting image embeddings) makes it a better choice to alternatives like t-SNE [19] for our application.

We generate 1D, 2D, and 3D projections of time embeddings $\mathbf{T}_{y_{\min},\ldots,y_{\max}}$ to analyse temporal structures and assess the chronological ordering by comparing the sequence of the 1D projections with the respective ground truth sequence.

To further investigate the *alignment* of temporal information across modalities (RQ2), we project image embeddings (of time-annotated images) into the low-dimensional subspaces that we previously generated for the time embeddings with the above dimensionality reduction techniques. If temporal information is aligned across the modalities, projections of image and time embeddings should be close in space.

3.3 Chronological Timeline Modelling (RQ3)

Time probing (see Section 3.1) treats each year independently, ignoring chronological relationships between temporal embeddings and failing to exploit temporal structure. The intuition behind our proposed approaches is that if temporal information forms a low-dimensional manifold in the embedding space, as our analysis suggests, we can exploit this structure to create an explicit "timeline" representation. Whether this representation maintains coherent chronological ordering is an open question that we aim to answer.

We propose two approaches for deriving a sequential time representation from time embeddings: (i) an implicit approach based on UMAP combined with hyperparameter optimisation and (ii) an explicit approach based on Bézier curve approximation. Both approaches offer alternatives to direct probing, with the Bézier curve [53] explicitly modeling temporal progression in a smooth and coherent manner and UMAP leveraging its ability to unfold non-linear manifolds [20] and being guided to organise time embeddings in a chronological order. Note that we just require the order of time embeddings to be chronological but not their sequence to be linear, i.e., equally spaced. We refer to such time representations in the following as "timelines".

3.3.1 UMAP-Based Timeline Representation. We apply UMAP to project the high-dimensional time embeddings $\mathbf{T}_y, y \in Y$ onto a one-dimensional timeline, to unfold the non-linear manifold of time embeddings. This process consists of three key steps:

1. Optimising UMAP parameters: UMAP has a number of hyperparameters (e.g., *n_neighbors* controlling local neighborhood size, *min_dist* controlling point separation in the embedding, and *metric* specifying the distance measure used). It is a priori not clear which parameters best unfold the manifold of time embeddings. Thus, we perform parameter optimisation to learn a transformation f_{UMAP} that best maps time embeddings $\mathbf{T}_y$ to their one-dimensional projections $\hat{\mathbf{T}}_y$. Thereby, UMAP parameters are optimised using the Tree-structured Parzen Estimator (TPE) method [36], which maximizes Spearman's rank correlation [46] between the projected order and the correct chronological order of years.

The transformation f_{UMAP} maps high-dimensional time embeddings $\mathbf{T}_y$ to a one-dimensional timeline as follows:

$$f_{\text{UMAP}} : \mathbf{R}^N \rightarrow \mathbf{R}^1, \quad \hat{\mathbf{T}}_y = f_{\text{UMAP}}(\mathbf{T}_y), \quad y \in Y,$$

where $\mathbf{T}_y \in \mathbf{R}^N$ represents the time embedding for year y in the VLM's embedding space of dimension N, and $\hat{\mathbf{T}}_y \in \mathbf{R}^1$ is its one-dimensional projection. TPE optimises the UMAP parameters by maximizing the following objective function: $\max \rho(Y, \hat{Y})$, where ρ denotes Spearman's rank correlation [46], which quantifies the monotonic relationship between the chronological order of years Y and their projected order $\hat{Y}$.

2. Projecting image embeddings: The learned transformation f_{UMAP} is applied to image embeddings $\mathbf{I}$, yielding their corresponding position on the timeline $\hat{\mathbf{I}}$.

3. Temporal inference: To retrieve the most likely year for an image, we identify the closest time embedding to its projection $\hat{\mathbf{I}}$ in the one dimensional projection space using Euclidean distance.

3.3.2 Bézier Curve-Based Timeline Representation. Alternatively to learning a suitable projection (which is time-consuming), we propose to model time via Bézier curve approximation in the original embedding space or a subspace of it. The Bézier curve approach provides a structured representation of time that enforces chronological consistency while capturing non-linear temporal relationships. Unlike time probing, which treats each year independently, and UMAP, which learns an implicit mapping, the Bézier curve explicitly models temporal progression, offering a transparent and inherently interpretable approach.

In a first step, we approximate the chronological sequence of time embeddings ($\mathbf{T}$) with a one-dimensional Bézier curve $C(t)$ that is embedded in $\mathbf{R}^N$, the embedding space of $\mathbf{T}$ (or a projection space $\mathbf{R}^S$ with dimension $S < N$ to reduce potential noise in temporal modeling and reduce computational complexity). The curve is fitted to all time embeddings from $y_{\min}$ to $y_{\max}$ and is discretized with $N_{samples}$, i.e., $t = 1, .., N_{samples}$. We leverage the *de Casteljau algorithm* [5] to interpolate the curve between control points $\mathcal{P}$ uniformly sampled from the sorted time embeddings. The number of control points $\mathcal{K}$ trades off over- vs. underfitting. Each embedding $\mathbf{T}_y$ is projected onto its closest point on $C(t)$ by:

$$\hat{\mathbf{T}}_y = \underset{t \in [0,1]}{\arg\min}\, d(\mathbf{T}_y, C(t)), \tag{2}$$

where $d(.,.)$ denotes Euclidean distance. To align image embeddings with the timeline, they are projected into the same space as the Bézier curve. This can be applied in the original embedding space or in a derived subspace of any dimension S (as long as the projection function g is known, i.e. $\hat{\mathbf{I}} = g(\mathbf{I})$ where $g : \mathbf{R}^N \rightarrow \mathbf{R}^S$).

For temporal inference, we propose two strategies: (i) nearest-neighbor matching, which assigns the closest projected time embedding, and (ii) interpolation between adjacent time points, i.e.:

$$y_{\text{pred}} = (1 - w)\, y_{\text{before}} + w y_{\text{after}}, \quad w = \frac{d(\hat{\mathbf{I}}, \hat{\mathbf{T}}_{\text{before}})}{d(\hat{\mathbf{I}}, \hat{\mathbf{T}}_{\text{before}}) + d(\hat{\mathbf{I}}, \hat{\mathbf{T}}_{\text{after}})}, \tag{3}$$

where $\hat{\mathbf{T}}_{\text{before}}$ and $\hat{\mathbf{T}}_{\text{after}}$ are the two closest projected time embeddings before and after $\hat{\mathbf{I}}$, and $y_{\text{before}}, y_{\text{after}}$ are their corresponding

years. Since interpolation yields fractional year values, this method allows for finer temporal granularity.

4 Dataset Curation

To evaluate the VLMs awareness to time, we compiled **TIME10k**, a dataset of over 10,000 time-annotated images showing human-made objects spanning a time range from 1715 to 2024 and six object categories: *Aircraft*, *Cars*, *Instruments*, *Mobile phones*, *Ships*, and *Weapons & Ammunition*. TIME10k was curated using Wikipedia [45] and Wikimedia Commons [44], leveraging Wikipedia's category system for object introduction dates. We extracted time-tagged objects and retrieved corresponding images, with manual verification to ensure quality. We adopted year-level granularity for temporal annotations as this yields the most consistent temporal categorization of the data and in most cases only years were available. Wikipedia's category system predominantly organises objects by their year of introduction (e.g., "Cars introduced in 2022"), and does not consistently provide finer-grained information. Overall, we consider a granularity of years a good tradeoff for evaluating a VLM's ability to position objects in time.

The dataset contains 10,091 images with varying sample numbers across classes (Cars: 4,393; Mobile Phones: 4,337; Ships: 841; Instruments: 436; Aircraft: 69; Weapons: 15). These object categories were selected based on the availability of time annotations in Wikipedia's hierarchical category system, ensuring reliable temporal annotations. The dataset shows a natural temporal bias toward more recent decades, with pre-1830s images primarily being drawings rather than photographs. This imbalance reflects historical developments and does not limit our evaluation approach. Furthermore, different object classes cover different time ranges, which can also be considered a real-world bias.

5 Experimental Setup

Here we detail the experimental setup for the evaluation of time-awareness in VLMs and to answer research questions RQ1-RQ3. We utilize TIME10k (Section 4) for all experiments, with images resized and normalised according to the specific VLM specifications.

5.1 Prompt Design

To examine the sensitivity of VLMs to linguistic variations in temporal queries, we evaluate multiple prompt formulations (P1-P9) that explore different ways of referencing the first appearance of an object. The prompts are detailed in Table 2. Additionally, we include minimalistic prompts (P1, P2) to assess the model's ability to infer time only from the indication of a year.

5.2 Vision Language Models

We evaluate a diverse selection of 37 VLMs with different architectures, backbones, and pretraining datasets. The models are applied in a zero-shot manner with frozen backbones (no fine-tuning). A detailed list of the selected VLMs, including the architectures and their chosen backbones is provided in Table 1.

5.3 Parameters

For *time probing* the pre-defined time range Y is $[Y_{\text{min}}, Y_{\text{max}}]$ where $Y_{\text{min}} = 1700$ and $Y_{\text{max}} = 2024$, resulting in $|Y| = 325$ individual years. No softmax was applied to the similarity scores. The evaluated text prompts are listed in detail in Table 2, and the evaluated VLMs in Table 1.

For *embedding space analysis*, UMAP parameters are optimised via TPE [36]. The optimisation yields the following parameters for the evaluated models: CLIP (ViT-B/32; $n_{\text{neighbors}} = 38$, min_dist = 0.7446) and EVA-CLIP (EVA02-CLIP-L-14-336; $n_{\text{neighbors}} = 21$ and min_dist = 0.1040). We consistently use cosine metric for UMAP.

For *timeline modeling*, we apply UMAP to obtain one-dimensional projections following the same optimisation procedure as in embedding space analysis. The Bézier curve is fitted using $\mathcal{K} = 200$ control points sampled uniformly along the time embeddings for the time range Y. We choose a relatively high $\mathcal{K}$ compared to the overall number of time embeddings of 325 to guarantee fidelity to the potentially non-linear shape of the time embeddings. For curve resolution we set $N_{\text{samples}} = 1000$ to allow for fine-grained interpolation along the curve.

5.4 Performance Metrics

We employ two types of metrics: ranking metrics to evaluate the quality of chronological orderings and accuracy metrics that assess time prediction quality. As ranking metrics we employ Spearman's Rank Correlation (ρ) [46], Kendall's Tau (τ) [38], and Modified Normalised Damerau-Levenshtein Distance (δ_{MNDL}) (adapted from [30]). δ_{MNDL} measures ranking errors based on adjacent swaps: $\delta_{\text{MNDL}} = 1 - 2\frac{S}{M}$, where S is the total number of adjacent swaps required to transform the predicted ranking into the correct order, and $M = \frac{N(N-1)}{2}$ is the maximum possible swaps for N elements. We adjust this metric to have the same range and meaning as ρ and τ. All metrics range from -1 to 1, where 1 indicates perfect chronological ordering and -1 indicates completely reversed ordering, both considered structurally correct as they maintain temporal relationships. These ranking metrics complement the following accuracy metrics by providing a comprehensive assessment of how well VLMs understand temporal relationships and chronological ordering, beyond just predicting specific years.

For accuracy evaluation, we use Mean Absolute Error (MAE) which measures the average deviation between predicted and ground truth years. Furthermore, we introduce TAI as a time-adaptive accuracy measure that applies greater tolerance for older years (reflecting inherent uncertainty in historical documentation) and stricter tolerance for more recent years. TAI is defined by a tolerance threshold function $T(y)$ and an intolerance threshold function $I(y)$ and performs linear interpolation between them. For a given image, TAI is computed as follows:

$$\text{TAI}_i = \begin{cases} 1, & \text{if } |y_{\text{pred}} - y_{\text{GT}}| \le T(y), \\ 1 - \frac{|y_{\text{pred}} - y_{\text{GT}}|}{I(y)}, & \text{if } T(y) < |y_{\text{pred}} - y_{\text{GT}}| < I(y), \\ 0, & \text{if } |y_{\text{pred}} - y_{\text{GT}}| \ge I(y). \end{cases} \quad (4)$$

Both thresholds vary over time. Their start and end values are explicitly defined (in years). In our case we set: $T_{Y_{\text{min}}} = 20$, $I_{Y_{\text{min}}} = 50$, $T_{Y_{\text{max}}} = 5$, $I_{Y_{\text{max}}} = 15$. These setting provide 20-year tolerance for historical versus 5-year for modern objects, which we consider a reasonable tradeoff for our investigation.

5.5 Evaluation Framework

Our evaluation consists of three main experiments: time probing, embedding space analysis, and chronological timeline modelling.

Time Probing (RQ1). To assess temporal awareness, we perform time probing on 37 VLMs (Table 1) using different text prompts (see Table 2). We further analyse the prompt sensitivity (Section 3.1) to examine its impact on predictions and scores.

Embedding Space Analysis (RQ2). To investigate the spatial structure of time embeddings, we apply dimensionality reduction techniques (Section 3.2) and evaluate their chronological alignment with the ground truth using ranking metrics. We further examine how image embeddings align with time embeddings to assess cross-modal consistency.

Chronological Timeline Modelling (RQ2 & RQ3). We evaluate timeline modelling using two representative VLMs: ViT-B/32 (CLIP) as the most popular contrastive model and EVA02-CLIP-B/16 (EVA-CLIP) for its strong performance (see Table 1). This focused selection serves as a proof-of-concept and reduces computational costs. Timeline quality is assessed via ranking metrics for chronological alignment (RQ2) and accuracy metrics (TAI, MAE) for time prediction (RQ3).

UMAP-Based Timeline Representation. UMAP is optimised to preserve temporal order and is then applied to image embeddings. Predictions are made by mapping images to the closest year. UMAP hyperparameters are detailed in Section 5.3.

Bézier Curve-Based Timeline Representation. We evaluate Bézier timeline fitting in both high-dimensional and KPCA reduced subspaces. Time inference is performed using nearest neighbour and interpolation (see Section 3.3.2). For visualisation, we reduce embeddings to $\mathbf{R}^3$ before curve fitting.

6 Results

We first present results for time probing on a broad set of VLMs (Section 6.1). Next, we present our analysis of temporal structures in the embedding space in Section 6.2 and finally, we demonstrate the capabilities of VLMs for timeline modeling (Section 6.3).

6.1 Time Probing

Quantitative results: Table 1 presents the MAE and TAI of 37 VLMs using prompt P7 (see Table 2) on the TIME10k dataset. OpenCLIP ("ViT-bigG-14-quickgelu") and EVA-CLIP ("EVA02-CLIP-L-14-336") achieve the best performance, followed by ImageBind [34] and ViT-Lens [42]. Overall, model architecture and training dataset have a strong influence on results. Some models fail at time modeling, possibly due to insufficient training datasets, e.g., "V1," "CommonPool," and "DFN2B" underperform, while "openai" and "Merged-2B" excel. Backbone complexity also matters, e.g., "ViTamin-S" performs poorly compared to its larger "ViTamin-XL-384" counterpart.

Prompt Sensitivity Analysis: Table 2 shows how different linguistic formulations affect time awareness performance across CLIP and EVA-CLIP models. The analysis demonstrates that prompt formulation is crucial for temporal prediction, with minimalistic prompts (P1, P2) performing significantly worse than more descriptive formulations.

Table 1: Evaluation of the awareness of time (in terms of MAE and TAI) across different VLMs with varying network architectures and backbones, pretrained on different data.

Architecture	Backbone	Training Data	↓MAE	↑ TAI
CLIP [3]	RN101	openai	10.61	0.70
CLIP [3]	RN50	openai	10.16	0.71
CLIP [3]	RN50x16	openai	9.40	0.73
CLIP [3]	RN50x4	openai	10.61	0.72
CLIP [3]	RN50x64	openai	7.26	0.81
CLIP [3]	ViT-B/16	openai	7.75	0.79
CLIP [3]	ViT-B/32	openai	9.53	0.70
CLIP [3]	ViT-L/14	openai	6.99	0.83
CLIP [3]	ViT-L/14@336px	openai	**6.76**	**0.84**
EVA-CLIP [31]	EVA01-CLIP-g-14	LAION-400M	8.87	0.78
EVA-CLIP [31]	EVA01-CLIP-g-14-plus	Merged-2B	11.00	0.80
EVA-CLIP [31]	EVA02-CLIP-B-16	Merged-2B	7.67	0.82
EVA-CLIP [31]	EVA02-CLIP-L-14	Merged-2B	6.30	0.86
EVA-CLIP [31]	EVA02-CLIP-L-14-336	Merged-2B	**6.20**	**0.86**
EVA-CLIP-18B [33]	EVA-CLIP-18B	Merged-2B+	6.45	0.86
EVA-CLIP-18B [33]	EVA-CLIP-8B	Merged-2B	6.53	0.84
EVA-CLIP-18B [33]	EVA-CLIP-8B-plus	Merged-2B	6.48	0.85
ImageBind [34]	ViT-H	LAION-2B	7.16	0.83
CoCa [51]	coca-VIT-L-14 [16]	MSCOCO+LAION2B	8.49	0.78
MobileCLIP [40]	MobileCLIP-S1	DataComp-DR	9.70	0.75
ViTamin [18]	ViTamin-S [16]	DataComp-1B	75.80	0.48
ViTamin [18]	ViTamin-XL-384 [16]	DataComp-1B	6.48	**0.86**
OpenCLIP [29]	RN50-quickgelu [16]	YFCC15M	29.27	0.28
OpenCLIP [29]	ViT-B-16 [16]	MetaCLIP	18.74	0.63
OpenCLIP [29]	ViT-B-16-plus-240 [16]	LAION400M	13.45	0.72
OpenCLIP [29]	ViT-B-16-quickgelu [16]	DFN2B	34.42	0.46
OpenCLIP [29]	xlm-roberta-base-VIT-B-32 [16]	LAION5B	16.92	0.66
OpenCLIP [29]	ViT-B-32	CommonPool	144.54	0.08
OpenCLIP [29]	ViT-bigG-14 [16]	LAION2B	8.18	0.81
OpenCLIP [29]	ViT-bigG-14-quickgelu [16]	MetaCLIP	**6.31**	0.85
OpenCLIP [29]	ViT-g-14 [16]	LAION2B	8.62	0.80
OpenCLIP [29]	convnext-xxlarge [16]	LAION2B AugReg	6.71	0.83
CLIPA [25]	ViT-H-14-CLIPA-336	DataComp-1B	9.25	0.82
SigLIP [52]	ViT-L-16-SigLIP-384 [16]	WebLi	12.93	0.71
SigLIP [52]	ViT-SO400M-14-SigLIP-384 [16]	WebLi	7.04	0.84
SigLIP [52]	nillb-clip-large-siglip [16]	V1	96.46	0.32
ViT-Lens [42]	ViT-Lens-L	Mixed Sources	**7.78**	**0.85**

Table 2: Prompt Sensitivity for CLIP and EVA-CLIP VLMs

Sentence Base	CLIP		EVA-CLIP	
	↓MAE	↑TAI	↓MAE	↑TAI
P1 - [year]	48.28	0.70	48.14	0.73
P2 - Year [year]	19.89	0.82	16.34	0.84
P3 - Was released in the year [year]	17.86	0.83	11.30	0.88
P4 - Was invented in the year [year]	19.34	0.81	9.18	0.89
P5 - Was first introduced in the year [year]	15.84	0.82	7.70	0.89
P6 - Was created in the year [year]	19.06	0.77	10.30	0.88
P7 - Was built in the year [year]	**8.79**	**0.86**	**7.44**	**0.89**
P8 - First appeared in the year [year]	15.19	0.82	14.24	0.85
P9 - Emerged in the year [year]	17.87	0.80	14.81	0.86

Class-wise performance: To examine how temporal awareness varies across object categories, we analyse the performance of EVA02-CLIP-L-14-336, one of our best-performing models, across the six classes in TIME10k (Table 3). Cars and mobile phones, which have been photographed more often in recent periods, achieve excellent performance (MAE $\leq$ 2.65, TAI $\geq$ 0.95), while categories like music instruments and weapons, whose depictions often predate the era of photography and have less distinct times of first appearance cannot be positioned that accurately in time (MAE > 30, TAI < 0.5). These results reflect the natural evolution of visual documentation and the fact that the classes have different annual ranges (see last row of Table 3). Notably, these class-specific scores

Table 3: Class-specific awareness to time.

Metric	Aircraft	Cars	Mobile Phones	Music Instruments	Ships	Weapons & Ammo
MAE ↓	14.25	2.64	2.65	32.94	18.12	33.50
TAI ↑	0.76	0.95	0.95	0.24	0.76	0.49
Annual range	[1893,2017]	[1888,2024]	[1984,2024]	[1715,2009]	[1744,1999]	[1939,2003]

do not directly average to those in Table 1 due to differences in class cardinalities.

Validity of time probing: Time probing selects the year with the highest dot-product similarity among all candidates (see Equation 1). To assess the validity of this choice, we analyse the distribution of scores across probed years using two approaches with one of the best-performing model EVA02-CLIP-L-14-336. *First*, we examine how often the top-ranked predictions match the ground truth year (Figure 3, left). The highest-scoring year (rank 1) is most frequently correct, followed by rank 2 and so on, confirming that the score reflects the probability of a correct assignment well. *Second*, we analyse the relationship between the two top-scoring years and the ground truth (Figure 3, right). Both highest (blue) and second-highest (orange) scores align well with the true year (red diagonal). Some outliers exist, particularly for recent years, where either the first or second-highest prediction is incorrect. Overall, the score distribution suggests that time probing reliably identifies the most probable year of first appearance.

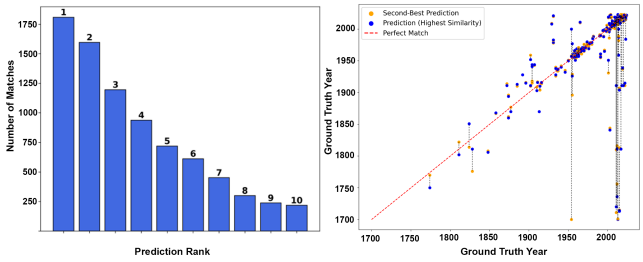

Figure 3: Left: The highest score (rank 1) is most often the correct year, followed by the second highest year (rank 2) etc. Right: Ground truth (red diagonal) vs. predicted year for highest (blue) and second-highest (orange) score. Vertical dotted lines represent their difference, which are mostly small.

6.2 Embedding Space Analysis

For embedding space analysis, we employ KPCA and UMAP and generate 1D, 2D and 3D projections of the time embeddings $\mathbf{T}_{1700}$,..., $\mathbf{T}_{2024}$ from the two selected representative models, ViT-B/32 (CLIP) and EVA02-CLIP-L-14-336 (EVA-CLIP). Figure 4 shows the example of a 3D projection of time embeddings from the ViT-B/32 (CLIP) model obtained using KPCA. As hypothesised, the projections align along a lower-dimensional (non-linear) manifold-like structure. Colour encodes time from 1700 (violet) to 2024 (yellow), revealing a chronological progression of time along this structure. Thus, RQ2 can be answered positively in this case. We observe a similar behavior for EVA-CLIP, indicating that this structure is consistent across different VLM architectures. Notably, this contrasts with recent findings in language models where temporal information seems to be encoded in linear subspaces [14, 15], highlighting that VLMs organise temporal information differently, namely in low-dimensional non-linear structures.

Table 4: Degree of chronological progression in 1D.

Metric	CLIP		EVA-CLIP	
	KPCA	UMAP	KPCA	UMAP
ρ	0.96	0.80	0.92	-0.70
τ	0.84	0.55	0.77	-0.48
δ_{MNDL}	0.84	0.55	0.77	0.74

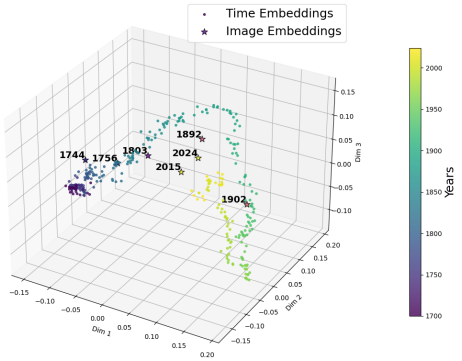

Figure 4: 3D KPCA projection of CLIP time and image embeddings.

To quantify chronological alignment, we generate 1D projections of $\mathbf{T}_{1700}$,..., $\mathbf{T}_{2024}$ using KPCA and UMAP with default parameters. Table 4 presents the respective ranking scores after reducing the embeddings to 1D and comparing the resulting order with the ground truth. The ranking metrics indicate that the temporal order is preserved to a large degree (e.g. Spearman correlation ρ of 0.96 for KPCA), showcasing a strong potential of the (projected) time embeddings to serve as a basis for deriving a timeline from it.

Finally, we project randomly selected image embeddings into the 3D space from Figure 4 (plotted as black stars with labels indicating their year). The image embeddings align well with the temporal structure, indicating temporal coherence across image and text modalities (RQ2).

6.3 Time Modeling and Prediction

We evaluate the two approaches for timeline modeling on ViT-B/32 (CLIP) and EVA02-CLIP-B/16 (EVA-CLIP). Figure 5 illustrates Bézier curve fitting in 3D projection space. Using $\mathcal{K} = 200$ chronologically sorted control points, we fit a 1D curve and map all time embeddings onto it to obtain a monotonic timeline. Figure 5(c) further shows image embeddings projected into the same space (black stars with year label), showing that the image embeddings align well with the curve.

Projection space dimension. For the Bézier approach the question arises, which projection space dimension is optimal, i.e. what is the minimal dimension necessary to capture the observed manifold-like structure. To answer this, we evaluate the Bézier approach for the entire TIME10k dataset for all possible projection dimensions

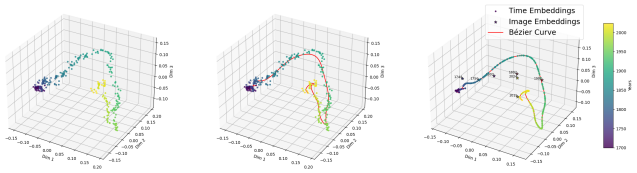

Figure 5: Fitting a Bézier curve to the time embeddings from CLIP in a 3D projection space. (a) Time embeddings coloured by year with selected control points marked with diamonds. (b) Fitting the Bézier curve. (c) Mapping points to the Bézier curve to obtain a 1D timeline representation.

and plot the MAE, see Figure 6 for results on two different VLMs. With increasing dimension, MAE reduces rapidly and then stays more or less constant. According to Figure 6, for both VLMs around 13 dimensions are already sufficient to capture most of the temporal information. Higher dimensions do not yield significant benefits. This finding is particularly interesting as it suggests that temporal information in VLMs is encoded in a remarkably compact subspace compared to the original embedding dimension (typically ≥ 512).

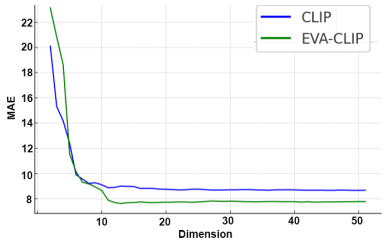

Figure 6: MAE per dimension for (EVA-)CLIP using KPCA.

Bezier approach variants. We evaluate four variants of the proposed Bézier-based approach for temporal inference that (i) vary in embedding space dimension and (ii) in the temporal inference method. Bézier($\mathbf{R}^N$,NN) and Bézier($\mathbf{R}^N$,Int) operate in the original N-dimensional embedding space and use either nearest neighbor ("NN") or linear interpolation ("Int") to assign the closest year, for potentially finer granularity. Bézier($\mathbf{R}^S$,NN) and Bézier($\mathbf{R}^S$,Int) operate on an S-dimensional subspace obtained via KPCA. Based on the result in Figure 6, we set $S = 13$.

Chronological structure. Table 5 compares time probing, UMAP, and the four Bézier-based methods for time prediction. Ranking scores (ρ, τ, δ_{MNDL}) between 0.96 and 0.99 confirm almost perfect chronological time representations. Note that for UMAP and EVA-CLIP the estimated timeline is inverted (negative correlation).

Performance comparison. The accuracy metrics (MAE, TAI) in Table 5 show that for CLIP the Bézier approaches clearly outperform UMAP and Time probing. For EVA-CLIP the Bézier approaches in average outperform UMAP and time probing in terms of TAI. For MAE the Bézier approach (with dimensionality reduction) yields a similar performance level to time probing. Interpolation does not significantly affect overall accuracy compared to nearest neighbor matching. Dimensionality reduction by KPCA consistently improves the Bézier approach, indicating that reducing dimensionality also removes unwanted factors or noise.

Table 5: Time prediction results for UMAP- and Bézier-based timeline extraction compared to time probing (baseline). Results outperforming time probing are highlighted in bold.

Metric	Time Probing	UMAP	Bézier($\mathbf{R}^n$,NN)	Bézier($\mathbf{R}^n$,Int)	Bézier($\mathbf{R}^s$,NN)	Bézier($\mathbf{R}^s$,Int)
CLIP						
MAE ↓	9.53	11.51	**9.10**	**9.16**	**9.00**	**8.81**
TAI ↑	0.70	**0.81**	**0.76**	**0.76**	**0.77**	**0.79**
Inference Time (ms) ↓	**5**	539	26	17	11	11
ρ	–	0.99	0.99	0.99	0.99	0.99
τ	–	0.98	0.99	0.99	0.99	0.99
δ_{MNDL}	–	0.98	0.99	0.99	0.99	0.99
EVA-CLIP						
MAE ↓	**7.67**	15.49	9.58	9.61	7.76	7.77
TAI ↑	0.82	0.74	**0.85**	0.72	**0.84**	**0.84**
Inference Time (ms) ↓	10	554	**9**	**9**	12	11
ρ	–	-0.99	0.99	0.99	0.99	0.99
τ	–	-0.96	0.98	0.98	0.98	0.98
δ_{MNDL}	–	-0.96	0.98	0.98	0.98	0.98

Runtime comparison. Experiments were conducted on an NVIDIA Tesla V100 GPU (32GB VRAM) with an AMD EPYC 9334 32-core processor and 64GB system memory. Inference times differ substantially between methods (see Table 5). The UMAP approach requires significantly more time (539-554ms per prediction) due to the projection process, while Bézier-based approaches achieve superior efficiency at 9-26ms and thus reach a competitive speed to time probing with the added benefit of getting an explicit timeline representation that can be efficiently re-used for time reasoning.

7 Conclusion

We presented a first systematic investigation of temporal awareness in vision-language models. We systematically evaluated 37 state-of-the-art VLMs with a new benchmark dataset (TIME10k) and found that modern models possess substantial temporal awareness, although performance varies significantly with training data quality and backbone model complexity. Most remarkably, we discovered that temporal information is encoded in a compact, low-dimensional manifold with strong chronological structure within the high-dimensional embedding space and developed novel approaches for the extraction of an explicit timeline representation based on UMAP and Beziér curve approximation. Still, there are certain limitations of our investigation motivating future work: (i) the effect of class imbalances and annual ranges needs to be further investigated; (ii) TIME10k in future should be extended with classes representing human-centric aspects (e.g. clothing); (iii) the evaluation of timeline extraction methods should be extended to more VLMs as well as generative VLMs to see if they exhibit similar temporal awareness (which we assume as they base on similar encoders). Ultimately, our work raises a fundamental question: if VLMs can implicitly capture and structure temporal information so effectively, can this approach generalize to other ordinal problems?

Acknowledgements

This work was funded by the Austrian Science Fund (FWF) under the Visual Heritage PhD Program doctoral grant [10.55776/DFH37].

References

[1] A. Nicolson, L. Schut, J. A. Noble, and Yarin Gal. 2024. Explaining explainability: understanding concept activation vectors. *arXiv.org*. doi:10.48550/arxiv.2404.03713.

[2] Alberto García-González, Alberto García-González, Antonio Huerta, Antonio Huerta, Sergio Zlotnik, Sergio Zlotnik, Pedro Díez, and Pedro Díez. 2020. A kernel principal component analysis (kPCA) digest with a new backward mapping (pre-image reconstruction) strategy. *arXiv: Numerical Analysis*, (Dec. 15, 2020). doi:10.21203/rs.3.rs-126052/v1.

[3] Alec Radford et al. 2021. Learning transferable visual models from natural language supervision. *International Conference on Machine Learning*.

[4] Alexandra Barancová, Melvin Wevers, and Nanne van Noord. 2023. Blind dates: examining the expression of temporality in historical photographs. *Workshop on Computational Humanities Research*, (Oct. 10, 2023). doi:10.48550/arxiv.2310.06633.

[5] Wolfgang Boehm and Andreas Müller. 1999. On de casteljau's algorithm. *Computer Aided Geometric Design*, 16, 7, 587–605.

[6] Xi Chen et al. 2023. Pali-3 vision language models: smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*.

[7] Xi Chen et al. 2022. Pali: a jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*.

[8] Cherag Aroraa, Tracy Holloway King, Jayant Kumar, Yi Lu, Sanat Sharma, Arvind Srikantan, David Uvalle, Josep Valls-Vargas, and Harsha Vardhan. 2024. Smart multi-modal search: contextual sparse and dense embedding integration in adobe express. *MMSR@CIKM*. doi:10.48550/arxiv.2408.14698.

[9] Christoph Schuhmann, Richard Vencu, Romain Beaumont, R. Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, J. Jitsev, and Aran Komatsuzaki. 2021. LAION-400m: open dataset of CLIP-filtered 400 million image-text pairs. *arXiv.org*.

[10] Duolikun Danier, Mehmet Aygün, Changjian Li, Hakan Bilen, and Oisin Mac Aodha. 2024. Depthcues: evaluating monocular depth perception in large vision models. *arXiv preprint arXiv:2411.17385*.

[11] Eric Ming Chen, Jin Sun, Apoorv Khandelwal, Dani Lischinski, Noah Snavely, and Hadar Averbuch-Elor. 2023. What's in a decade? transforming faces through time. *Computer graphics forum (Print)*, 42, 2, (May 1, 2023), 281–291. doi:10.1111/cgf.14761.

[12] Erik Härkönen, Erik Härkönen, Aaron Hertzmann, Aaron Hertzmann, Jaakko Lehtinen, Jaakko Lehtinen, Sylvain Paris, and Sylvain Paris. 2020. GANSpace: discovering interpretable GAN controls. *Neural Information Processing Systems*, 33, 9841–9850.

[13] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander Toshev, and Vaishaal Shankar. 2023. Data filtering networks. *arXiv preprint arXiv:2309.17425*.

[14] Wes Gurnee and Max Tegmark. 2023. Language models represent space and time. *arXiv preprint arXiv:2310.02207*.

[15] Benjamin Heinzerling and Kentaro Inui. 2024. Monotonic representation of numeric properties in language models. *arXiv preprint arXiv:2403.10381*.

[16] [SW] Gabriel Ilharco et al., OpenCLIP version 0.1, July 2021. doi:10.5281/zenodo.5143773, URL: https://doi.org/10.5281/zenodo.5143773.

[17] Jiazhi Xia et al. 2022. Interactive visual cluster analysis by contrastive dimensionality reduction. *IEEE Transactions on Visualization and Computer Graphics*, PP, (Sept. 27, 2022). doi:10.1109/tvcg.2022.3209423.

[18] Jieneng Chen, Qihang Yu, Xiaohui Shen, Alan L. Yuille, and Liang-Chieh Chen. 2024. ViTamin: designing scalable vision models in the vision-language era. *Computer Vision and Pattern Recognition*. doi:10.1109/cvpr52733.2024.01231.

[19] Laurens van der Maaten, Laurens van der Maaten, Geoffrey E. Hinton, and Geoffrey E. Hinton. 2008. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9, 86, (Jan. 1, 2008), 2579–2605.

[20] Leland McInnes, Leland McInnes, John J. Healy, and John Healy. 2018. UMAP: uniform manifold approximation and projection for dimension reduction. *arXiv: Machine Learning*, (Feb. 9, 2018).

[21] Paul Lerner, Olivier Ferret, and Camille Guinaudeau. 2024. Cross-modal retrieval for knowledge-based visual question answering. In *European Conference on Information Retrieval*. Springer, 421–438.

[22] Martha Lewis, Nihal V Nayak, Peilin Yu, Qinan Yu, Jack Merullo, Stephen H Bach, and Ellie Pavlick. 2022. Does clip bind concepts? probing compositionality in large image models. *arXiv preprint arXiv:2212.10537*.

[23] Jeffrey Li et al. 2024. Datacomp-lm: in search of the next generation of training sets for language models. *Advances in Neural Information Processing Systems*, 37, 14200–14282.

[24] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*. PMLR, 19730–19742.

[25] Xianhang Li, Zeyu Wang, and Cihang Xie. 2024. An inverse scaling law for clip training. *Advances in Neural Information Processing Systems*, 36.

[26] Qiaoling Lin, Shuang Wang, Xiutiao Ye, Ruixuan Wang, Rui Yang, and Licheng Jiao. 2024. Clip-based grid features and masking for remote sensing image captioning. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*.

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36, 34892–34916.

[28] Leland McInnes, John Healy, and James Melville. 2018. Umap: uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

[29] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. *Computer Vision and Pattern Recognition*, (June 1, 2023). doi:10.1109/cvpr52729.2023.00276.

[30] Frederic P Miller, Agnes F Vandome, and John McBrewster. 2009. Levenshtein distance: information theory, computer science, string (computer science), string metric, damerau? levenshtein distance, spell checker, hamming distance. (2009).

[31] Qiang Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yong Cao. 2023. EVA-CLIP: improved training techniques for CLIP at scale. *arXiv.org*, (Mar. 27, 2023). doi:10.48550/arxiv.2303.15389.

[32] Qizhou Wang, Yong Lin, Yongqiang Chen, Ludwig Schmidt, Bo Han, and Tong Zhang. 2024. A sober look at the robustness of CLIPs to spurious features.

[33] Quan Sun, Jinsheng Wang, Qiying Yu, Yufeng Cui, Fan Zhang, Xiaosong Zhang, and Xinlong Wang. 2024. EVA-CLIP-18b: scaling CLIP to 18 billion parameters. *arXiv.org*. doi:10.48550/arxiv.2402.04252.

[34] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. 2023. ImageBind one embedding space to bind them all. *Computer Vision and Pattern Recognition*, (June 1, 2023). doi:10.1109/cvpr52729.2023.01457.

[35] Schuhmann, Christoph et al. 2022. LAION-5b: an open large-scale dataset for training next generation image-text models. *Cornell University - arXiv*, (Oct. 15, 2022). doi:10.48550/arxiv.2210.08402.

[36] Shuhei Watanabe. 2023. Tree-structured parzen estimator: understanding its algorithm components and their roles for better empirical performance. *arXiv.org*, (Apr. 21, 2023). doi:10.48550/arxiv.2304.11127.

[37] Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th international ACM SIGIR conference on research and development in information retrieval*, 2443–2449.

[38] Alexei Stepanov. 2015. On the kendall correlation coefficient. *arXiv preprint arXiv:1507.01427*.

[39] Ben Vardi, Oron Nir, and Ariel Shamir. 2025. Clip-up: clip-based unanswerable problem detection for visual question answering. *arXiv preprint arXiv:2501.01371*.

[40] Pavan Kumar Anasosalu Vasu, Hadi Pouransari, Fartash Faghri, Raviteja Vemulapalli, and Oncel Tuzel. 2024. Mobileclip: fast image-text models through multi-modal reinforced training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15963–15974.

[41] Wang, Jianyi, Jianyi Wang, Chan, Kelvin C. K., Kelvin C. K. Chan, Loy, Chen Change, and Chen Change Loy. 2022. Exploring CLIP for assessing the look and feel of images. *AAAI Conference on Artificial Intelligence*, (July 25, 2022). doi:10.48550/arxiv.2207.12396.

[42] Weixian Lei, Yixiao Ge, Jianfeng Zhang, Dylan Sun, Kun Yi, Ying Shan, and Mike Zheng Shou. 2023. VIT-LENS: towards omni-modal representations. *Computer Vision and Pattern Recognition*. doi:10.1109/cvpr52733.2024.02516.

[43] Melvin Wevers. 2021. Scene detection in de boer historical photo collection. In *Proceedings of the 13th International Conference on Agents and Artificial Intelligence - Volume 1: ARTIDIGH*. INSTICC. SciTePress, 601–610. ISBN: 978-989-758-484-8. doi:10.5220/0010288206010610.

[44] Wikimedia Foundation. 2025. Wikimedia Foundation. (2025). https://www.wikimedia.org.

[45] Wikimedia Foundation. 2025. Wikipedia — the free encyclopedia. (2025). https://www.wikipedia.org.

[46] Clark Wissler. 1905. The spearman correlation formula. *Science*, 22, 558, 309–311.

[47] Xi Ding and Lei Wang. 2024. Do language models understand time?

[48] Xiao Dong, Runhui Huang, Xiaoyong Wei, Zequn Jie, Jianxing Yu, Jian Yin, and Xiaodan Liang. 2023. UniDiff: advancing vision-language models with generative and discriminative learning. *arXiv.org*, (June 1, 2023). doi:10.48550/arxiv.2306.00813.

[49] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *IEEE International Conference on Computer Vision*. doi:10.1109/iccv51070.2023.01100.

[50] Hu Xu, Saining Xie, Po-Yao Huang, Licheng Yu, Russell Howes, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. 2023. Cit: curation in training for effective vision-language data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15180–15189.

[51] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

[52] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, 11975–11986.

[53] Linghui Zhao, Jingqing Jiang, Chuyi Song, Lanying Bao, and Jingying Gao. 2013. Parameter optimization for bezier curve fitting based on genetic algorithm. In *Advances in Swarm Intelligence: 4th International Conference, ICSI 2013, Harbin, China, June 12-15, 2013, Proceedings, Part I 4*. Springer, 451–458.

[54] Ziyu Wan et al. 2020. Bringing old photos back to life. *Computer Vision and Pattern Recognition*, (June 14, 2020), 2747–2757. doi:10.1109/cvpr42600.2020.00282.