VDBENCH: A BENCHMARK FOR VISUAL DIACHRONIC ANALYSIS OF HISTORICAL DEPICTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce *Visual Diachronic Analysis* (VDA), a new computer vision challenge for studying how the visual context of a depicted subject or theme changes over time and how these changes relate to documented historical shifts in meaning. This raises the question to what extent such changes in meaning can be inferred computationally from historical depictions. To study such phenomena, we introduce VDBench, the first benchmark for VDA. In the absence of available and sufficiently annotated historical image collections, we propose an unconventional approach: we construct a historically grounded synthetic benchmark dataset. We compile a source-grounded collection of human-centred diachronic topics and design a controlled text-to-image pipeline to generate plausible synthetic historical depictions. Domain experts confirm the suitability and plausibility of the data. We specify several computer vision tasks, propose a suitable evaluation methodology, and provide baseline methods and implementations as reference for future work. The dataset and an interactive explorer are available at https://anonymousauthor287.github.io/projectpage/. All code will be released upon acceptance.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

1 INTRODUCTION

In this paper we introduce *visual diachronic analysis* as a novel computer vision task. We present a benchmark dataset, formulate the corresponding analysis tasks, specify an evaluation framework, and provide first baselines.

Diachronic analysis has its origins in linguistics and examines how language evolves over time (Saussure, 1959). *Historical semantics* studies a specific form of this change, i.e. shifts in a word's meaning while its lexical form remains stable (Traugott & Dasher, 2002). Such shifts were traditionally studied through close philological readings of historical texts. Over the past decade, computational methods have enabled their systematic study through lexical semantic change detection (Tahmasebi et al., 2021). Hamilton et al. (2016), for example, align word embeddings trained on different historical periods. For the word "gay", its nearest neighbours in the embedding space, which represent words used in similar contexts, shift from "cheerful" and "frolicsome" towards "homosexual" and "lesbian" across

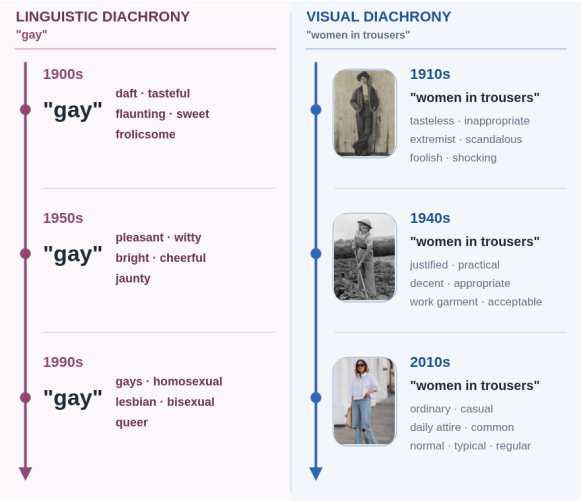

Figure 1: *Linguistic* and *visual diachrony*. Left: "gay" shifts from denoting cheerfulness to homosexuality across the twentieth century, adapted from (Hamilton et al., 2016). Right: the social meaning of women wearing trousers shifts from scandalous in the 1910s to practical in the 1940s and ordinary in the 2010s, partly reflected in the visual context. VDA aims to identify such changes automatically from images.

the twentieth century (Fig. 1, left). Here, the semantic change is reflected in systematic shifts in the word's *context*, rather than in its lexical form.

From linguistic to visual diachronic analysis. The linguistic study of diachronic change motivates an analogous perspective in the visual domain. Similar to a word whose form stays stable while its usage shifts, a recognisable subject may persist across historical periods while its depicted context and associated meaning change. Figure 1 (right) illustrates this by the example of the topic "women in trousers". In the 1910s, women wearing trousers were associated with suffrage activism and challenges to prevailing dress conventions (Cunningham, 2003; Atkinson, 2018). Depictions from this period may therefore place the subject in public demonstrations alongside other activists or political contexts. By the 1940s, depictions increasingly place women wearing trousers in wartime workplaces, including factories and industrial production sites (Turner, 2021). This context reflects a different social meaning of the garment, with trousers becoming accepted as workwear. In later decades, their increasing appearance in everyday settings indicates broader acceptance in Western societies. The example shows how changes in the visual context surrounding a persistent subject can provide observable signs of historical and social meaning shifts. Therefore in visual diachronic analysis (VDA) we ask: *to what extent can computational methods automatically identify such cues in historical depictions and use them to infer changes in meaning over time?* Section 2 defines VDA, its related concepts, and terminology in detail.

Research gap. To date, no dataset, evaluation framework, or targeted method exists for VDA. Existing computer vision research addresses related components in isolation, including object, scene, and activity recognition, contextual reasoning (Galleguillos & Belongie, 2010; Wang & Zhu, 2023), compositional understanding in vision-language models (Yuksekgonul et al., 2023), contextual distribution shifts in image collections (Liang & Zou, 2022), temporal representations (Barancová et al., 2023; Tekaya et al., 2025), and change discovery across large image collections (Deng et al., 2025). However, these methods do not address VDA as a whole, and no benchmark evaluates whether visual depictions can recover historically documented changes in context and meaning.

Challenges. Such a benchmark is hard to construct, since it requires structured, historically grounded information about which subjects underwent semantic change, when these changes occurred, and which observable visual contexts are associated with them. Historical image archives typically do not organise or annotate their holdings by diachronic topics and thus cannot provide the required ground truth. Furthermore, a benchmark dataset should sufficiently sample all topics and periods, which is hard in practice due to the varying availability of historical images.

A benchmark for VDA. The sparse availability of data motivated us to pursue a rather unconventional approach for data compilation. *First*, we start from documented history rather than images. Instead of retrospectively identifying diachronic topics in image archives, we construct the benchmark from documented historical developments and phenomena. Our data compilation strategy leverages an LLM-based retrieval pipeline with human verification along the process. In this way, we semi-automatically compile a source-grounded collection of visual diachronic topics from historical literature, specifying their periods, semantic changes, and associated visual contexts, and validate them with humanities experts. *Second*, in the absence of sufficiently annotated and publicly available data, we propose a complex image-generation pipeline that generates plausible synthetic historical depictions for each documented period. The generated images represent the documented topics and provide controlled proxy ground truth for evaluating methods for VDA. Domain experts positively assessed image plausibility on a randomly selected subset.

The **contributions** of this paper are:

1. **The novel computer vision challenge of "visual diachronic analysis"** (VDA) that aims to discover historical changes in visual context and meaning from images, including two concrete benchmark tasks, i.e. *visual context detection* and *topic and trajectory discovery*.
2. **The VDBench benchmark**, including a first historically grounded **dataset** of visual diachronic topics with plausible synthetic depictions together with sophisticated semi-automatic **data retrieval and generation pipelines** and an **evaluation framework** for the specified tasks.
3. **First computational baselines** for the two proposed VDA tasks.

2 VISUAL DIACHRONIC ANALYSIS

Visual diachrony. We define VDA as the study of how the visual context associated with a recognisable subject changes across historical periods. A subject may be a person, object, activity, or another recurring visual element of interest. We call such a subject, together with the periods over which its contextual change is documented, a *visual diachronic topic*. The evolution of its visual context across these periods forms its *diachronic trajectory*. Without loss of generality, in this work we focus on human-centred topics that reflect questions studied in history and sociology. For example, sociologists have examined how smoking changed from a socially accepted practice to an increasingly stigmatised one under shifting policies and social norms (Stuber et al., 2008). Our topics include social practices, activities, clothing, and culturally significant objects whose changing use and representation can provide evidence of broader social and cultural change.

Visual and semantic context. Computer vision commonly distinguishes between *content*, the entities and attributes depicted in an image (Dubey, 2022), and *context*, the surrounding information that contributes to their interpretation (Jain & Sinha, 2010; Bazire & Brézillon, 2005). Context has been characterised along several dimensions, including temporal, spatial, and semantic context (Dey et al., 2001; Brézillon, 1999). Our formulation makes a central distinction between *visual context* and *semantic context*. Visual context comprises directly observable information, such as the subject's environment, depicted activity, and co-occurring visual elements. Semantic context captures the historically situated meanings, associations, and social significance connected to the subject. Although semantic context cannot generally be determined from an image alone, changes in visual context can indicate semantic shifts over time. We operationalise this formulation by adapting Dey et al.'s (2001) categories of identity, location (*where*), activity (*what*), and time (*when*): *identity* is the recurring subject, *where* is its environment, *what* is represented by its activity and other co-occurring objects in the scene, and *when* is the year the depiction originates from.

With these concepts in place, VDBench defines a number of diachronic topics from historical and sociological research together with their trajectories and respective depictions. The next section describes how we compiled this source-grounded topic collection and how we populated it with plausible images. The validation of the selected trajectories and generated depictions is presented in Sec. 4, followed by a definition of the benchmark tasks in Sec. 5.1 and first baseline results in Sec. 6.

3 VDBENCH DATASET

The VDBench dataset comprises (i) a source-grounded collection of historical trajectories and (ii) for each trajectory a collection of images synthetically generated under controlled conditions. Table 1 provides the dataset's metrics and Fig. 2 outlines its construction and validation.

Table 1: VDBench scale. Visual vocabularies are the predefined sets of environment, activity, and dated-trend labels used to specify the context of generated images.

Collection of trajectories		*Visual vocabularies*		*Collection of images*	
Topics (proposed / retained)	266 / 148	Environments	669	Topics	41
Periods (with descriptions)	706	Activities	177	Periods	201
Verified publications/webpages	174	Dated trends	74	Images, 100 per period	20,100

Source-grounded topic construction. To establish historically grounded trajectories before image generation, each topic is constructed from documented claims. Using separate role-specific prompts, DeepSeek-R1 (DeepSeek-AI, 2025) serves as the *Historian* and *Librarian* agents. The *Historian* proposes a photographable topic, identifies its defining periods, and describes their visible characteristics and historical meanings alongside claim-specific search terms. The *Librarian* evaluates the retrieved sources and links each source to the claims it supports. Their complete prompts are provided in App. C (Listings 1 and 2). Sources are primarily scholarly articles and books, supplemented by reference, institutional, and archival webpages, including Wikipedia. Bibliographic records are verified primarily through the Crossref API, with OpenAlex, Semantic Scholar, Open Library, and publisher or repository pages used when needed. Topic-level human validation occurs in two steps. First, the authors removed topics with invalid sources or (obviously) unsupported claims, and second, experts reviewed a selected subset of the remaining trajectories to assess their

appropriateness. Details are provided in Sec. 4. Of the 266 topics proposed by our pipeline, 148 were retained after source audit, yielding 706 period descriptions supported by 174 verified sources.

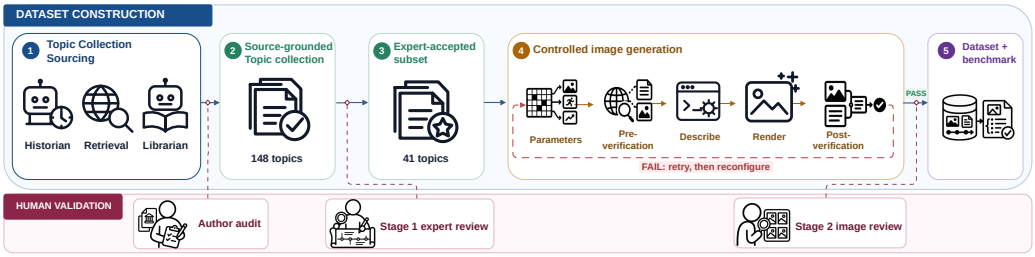

Figure 2: Overview of the semi-automatic VDBench dataset construction.

Controlled image generation. Images are generated only for the 41 topics approved by the experts in Stage 1 (Sec. 4). For each of their periods, the pipeline first constructs a structured description of visual context through the *when*, *where*, and *what* dimensions defined in Sec. 2. The topic fixes the depicted subject. *When* is a point year within the period; *where* is an environment selected from 669 scene classes compiled from Places365 (Zhou et al., 2018) and SUN (Xiao et al., 2010); and *what* comprises an activity selected from 177 time-use classes compiled from ATUS (U.S. Bureau of Labor Statistics, 2023), HETUS (Eurostat, 2020), and MTUS (Centre for Time Use Research, 2020), together with a caption describing the remaining intended visual context. Some specifications additionally include a *dated trend*, an optional photographable cultural phenomenon associated with a documented time and taken from a 74-item vocabulary compiled from dated Wikipedia categories (Wikipedia contributors, 2026).

Specifying diverse but controlled visual context. For each image, the pipeline begins with a topic and one of its periods and selects one environment, one activity, and, for some images, one dated trend. Varying these selections across the 100 images of a period diversifies the depictions, while the fixed topic and period keep them controlled. The topic specifies the subject, the environment *where*, and the activity the controlled part of *what*. The period bounds the later choice of *when*.

Grounding the selected context. To exclude implausible combinations before generation, DuckDuckGo Image Search (DuckDuckGo, 2026) retrieves candidate images for the proposed topic, environment, and activity. CLIP (Radford et al., 2021) uses zero-shot image-text matching to check the topic subject together with the selected environment and activity in the retrieved images. Only visually supported combinations proceed to image generation. The optional dated trend is added later to the generation prompt rather than included in this check, as requiring it in the retrieved images would overly restrict the possible combinations. Further details are provided in App. B.

Generate, verify, and record. For each combination that passes the grounding step, GPT-5.6 Sol (`gpt-5.6-sol`) (OpenAI, 2026) receives the period description and its start and end years, and chooses a year within this range as the depicted year. For each period, 5% of the generation prompts also include a dated trend. The model then produces a detailed image-generation prompt and a caption describing the intended visual content (the complete prompt is in App. C, Listing 3). Nano Banana Fast (Google DeepMind, 2025) renders the image. CLIP uses zero-shot image-text matching to check whether the selected environment, activity, and optional trend are visible. Images whose chosen elements are not detected are regenerated, with a new combination if needed.

Resulting image collection. The collection contains 20,100 images across 41 topics and 201 periods, with 100 images per period. Each released record contains its topic, period, year, environment, activity, and caption; any dated trend is retained only as construction metadata. Open-vocabulary methods use the complete collection without VDBench training. For supervised methods, the release provides two train, validation, and test schemes: a uniform split, in which every topic appears in all three sets, and a topic-disjoint split, which holds out complete topics to assess generalisation.

4 Dataset Validation

VDBench is validated in three steps during its construction. First, the authors review the proposed topics and remove those with unsupported claims or unverifiable sources. Second, domain experts

review a subset of the remaining trajectories before image generation in Stage 1 and third, experts assess a random sample of the generated images. Table 2 summarises both expert stages.

Stage 1: Topic and trajectory review. Experts in gender studies, sociology, social anthropology, literary studies, digital humanities, film archiving and theory, and migration history participated in the review. They could browse the complete topic list and selected only topics within their expertise. For each selected topic, they judged whether it showed meaningful visual change over time and represented a culturally important phenomenon. These judgements concerned the topic as a documented phenomenon rather than personal agreement with its social or political meanings. They then examined the complete sequence of periods and descriptions. They also rated each period description on a 1–5 agreement scale and could revise period boundaries (in years) and provide comments. Final confirmation of a topic followed a strict majority of completed reviews, excluding ties. Outcomes and the review interface are in App. D, Tab. 10, and Fig. 4.

Stage 2: Image plausibility review. Four of the six Stage 1 reviewers assessed images generated for topics within their expertise. For each period, they saw the same topic title, complete trajectory, and period descriptions used in Stage 1, together with a random sample of 3 of the 100 generated images. Reviewers were explicitly informed that the images were AI-generated. They were asked to judge their visual plausibility and correspondence to the stated topic and period rather than identify generation artifacts. Five topics received two reviews, covering 53 images assessed by both reviewers. The resulting agreement is reported in Tab. 2. Detailed outcomes and the review interface are provided in App. D, Tab. 11 and Fig. 5.

Table 2: Expert-validation results.

Author filtering	
Topics proposed	266
Topics retained	148 (55.6%)
Topics removed	118 (44.4%)
Stage 1: Trajectory review	
Experts	6
Topic reviews	57
Topics (>1 review)	9
Approved/rejected/tied	41/3/2
Period ratings	270
Periods (>1 rating)	41
Mean/median	4.02/4.00 (1–5)
Stage 2: Image plausibility	
Experts	4/6
Topic reviews	34
Topics (>1 review)	5
Period reviews	167
Periods (>1 review)	25
Image judgements	501
Plausible images	422/501 (84.2%)
Images (>1 review)	53
Same judgement	38/53 (71.7%)

Validation outcomes and limitations. Author filtering retained 148 of 266 proposed topics. Stage 1 experts approved 41 of 46 reviewed topics, rejected 3, and tied on 2. The 270 period ratings averaged 4.02/5, with a median of 4.00. In Stage 2, 422 of 501 judgements across 29 topics rated an image as plausible (84.2%). Per-reviewer plausibility rates ranged from 71.7% to 93.8% across their topic sets. Reviewers agreed on 38 of 53 images assessed by both (71.7%). Expert comments highlighted geographic scope, disputed period boundaries, and limited visual recognisability for some topics. These observations are summarised in App. D, Tab. 10. With the data validated, we turn to the tasks it supports.

5 Benchmark Tasks

5.1 Task Definition

We evaluate VDA through two complementary computer vision tasks that build upon each other. Task 1 estimates the temporal and visual context of an individual image, while Task 2 discovers a visual diachronic topic and its trajectory from a given image set with contextual metadata. The tasks can be evaluated separately or together. Task 2 receives image-based ground truth when evaluated separately or alternatively Task 1 predictions when the tasks are evaluated together. Figure 3 and Table 3 summarise their respective inputs and outputs.

Using the context dimensions defined in Sec. 2, T denotes a visual diachronic topic centred on one recurring subject. For image I_i, t_i denotes the year it was taken (*when*), e_i its spatial context, i.e. the environment (*where*), and (a_i, c_i) the activity shown and a caption describing the visual context (*what*). We represent a topic as $T = (n, d, \mathcal{I}, \mathcal{P})$, where n is its title, d a description of the diachronic topic, $\mathcal{I} = \{I_1, \ldots, I_N\}$ the set of images assigned to it, and $\mathcal{P} = (P_1, \ldots, P_J)$ an ordered list of periods defining the diachronic trajectory of the topic. Each period $P_j = (\mathcal{I}_j, s_j, R_j, m_j)$ comprises its assigned images $\mathcal{I}_j$, a free-text title s_j naming their shared observable visual context, a range of years R_j, and a free-text description m_j of its historically situated meaning. Each image belongs to

exactly one period, so $\mathcal{I}_1, \ldots, \mathcal{I}_J$ partition $\mathcal{I}$ into disjoint subsets. Period ranges still may overlap, as historical states can coexist during gradual transitions of periods. Indices i, j, and k denote images, ground-truth periods, and predicted periods, respectively. We indicate predicted variables with a hat.

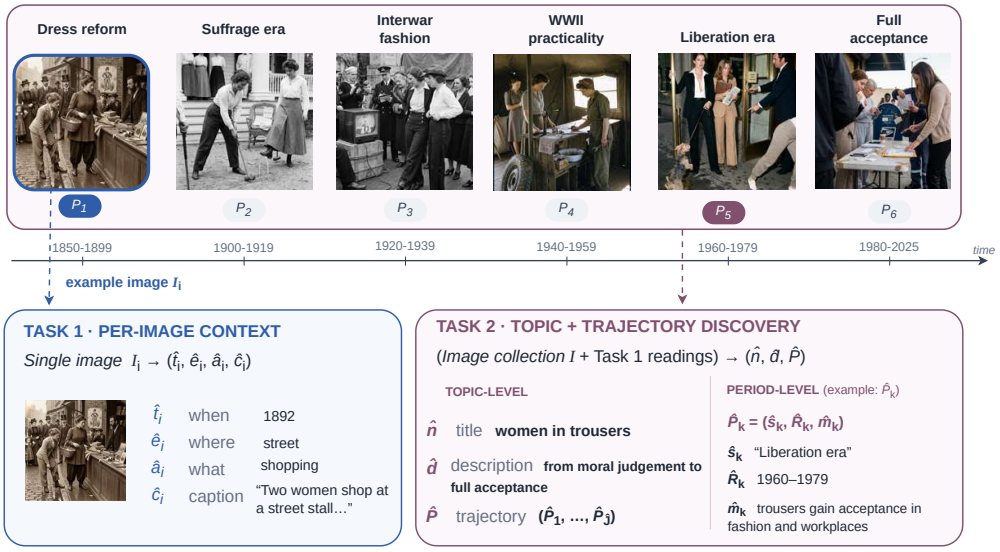

Figure 3: Illustration of the two benchmark tasks on the *women in trousers* example.

Table 3: Overview of the benchmark tasks.

Task 1: Visual context detection	
in: single image I_i	
out: $\langle \hat{t}_i, \hat{e}_i, \hat{a}_i, \hat{c}_i \rangle$	
$\hat{t}_i$ *when*	estimated year in which the image was taken
$\hat{e}_i$ *where*	environment (from a closed vocabulary, see Sec. 3)
$\hat{a}_i$ *what*	activity (from a closed vocabulary, see Sec. 3)
$\hat{c}_i$ caption	free-form description of the observable visual context
Task 2: Topic and trajectory discovery	
in: image set $\mathcal{I}$ for one topic and predictions or ground truth from Task 1 for each image	
out: topic title $\hat{n}$, overall description $\hat{d}$, period sequence $\hat{\mathcal{P}}$	
Title	$\hat{n}$, topic title
Description	$\hat{d}$, overall description of the diachronic topic
Sequence of periods	$\hat{\mathcal{P}} = (\hat{P}_1, \ldots, \hat{P}_{\hat{J}})$, chronologically ordered periods (year ranges can overlap)
Image assignments	$\hat{\mathcal{I}}_k \subseteq \mathcal{I}$, images assigned to period k
Period title	$\hat{s}_k$, title of period k
Range	$\hat{R}_k$, inferred year range of period k
Period description	$\hat{m}_k$, description capturing characteristics of the period and providing indications on the historical meaning of the k-th period

Task 1: Visual context detection. Given an image I_i, Task 1 predicts the year it was taken, the environment, activity, and context description as $I_i \mapsto (\hat{t}_i, \hat{e}_i, \hat{a}_i, \hat{c}_i)$. Environment and activity are drawn from the predefined vocabularies (Sec. 3), while context description c_i is open-vocabulary.

Task 2: Topic and trajectory discovery. Task 2 receives an image set $\mathcal{I}$ belonging to a single topic whose identity is hidden. A method must therefore infer the topic and express it through the predicted title $\hat{n}$ and description $\hat{d}$. When Task 2 is evaluated separately, it receives ground truth per image, i.e. (t_i, e_i, a_i, c_i). When the two tasks are evaluated together, it receives the corresponding Task 1 predictions $(\hat{t}_i, \hat{e}_i, \hat{a}_i, \hat{c}_i)$. No Task 2 ground truth or human input is provided in either case. The goal of Task 2 is to partition the overall trajectory into a suitable number of periods $\hat{J}$, assign every image to exactly one period, and describe the periods accordingly. When Task 1 predictions

are used, the range of period k is derived as $\hat{R}_k = [\min_{i \in \hat{\mathcal{I}}_k} \hat{t}_i, \max_{i \in \hat{\mathcal{I}}_k} \hat{t}_i]$. When ground-truth values are used, the same calculation uses t_i. Each predicted period $\hat{P}_k = (\hat{\mathcal{I}}_k, \hat{s}_k, \hat{R}_k, \hat{m}_k)$ contains its assigned images $\hat{\mathcal{I}}_k$, title $\hat{s}_k$, derived range $\hat{R}_k$, and description $\hat{m}_k$. The complete output is $(\hat{n}, \hat{d}, \hat{\mathcal{P}})$, where $\hat{\mathcal{P}} = (\hat{P}_1, \ldots, \hat{P}_{\hat{J}})$. Section 5.2 defines the evaluation protocol for both tasks.

5.2 Evaluation Protocol

Table 4 lists the metrics used for each benchmark output. Task 1 metrics are computed over all input images. Task 2 supports two settings. *Independent* evaluation (GT), where the tasks are evaluated separately, uses the ground-truth values from Task 1 (t_i, e_i, a_i, c_i), while end-to-end evaluation (Pred.), where they are evaluated together, uses the corresponding Task 1 predictions. The same Task 2 metrics are computed per topic and are macro-averaged so that every topic contributes equally. Each setting is reported separately using the full metric set rather than a single aggregation.

Per-image outputs. We evaluate year predictions using MAE in years, and environment and activity using top-1 accuracy and macro-F1. Predictions span all 669 environment and 177 activity classes; macro-F1 averages per-class F1 only over labels represented in the ground truth (see Sec. 3). The context description $\hat{c}_i$ uses the text metrics below.

Free-text outputs.

Context descriptions, topic and period titles, and topic and period descriptions are evaluated with BERTScore-F1 (Zhang et al., 2020) and MoverScore (Zhao et al., 2019). Both tolerate paraphrases: BERTScore aligns contextualised tokens, whereas MoverScore uses optimal transport over contextualised n-grams, providing complementary local and global comparisons.

Segmentations into periods and period count. These metrics assess whether images are assigned to the correct periods and whether the correct number of periods is recovered. We leverage cluster validation indices to compare ground-truth periods with predicted segmentations. ARI (Hubert & Arabie, 1985) assesses whether each pair of images is placed together or apart in the same way in the predicted and ground-truth groupings. AMI (Vinh et al., 2010) measures the information shared by the two groupings. Both provide complementary pair-counting and information-theoretic views of grouping agreement. The period-count error $|\hat{J} - J|$ is the absolute difference between the predicted and ground-truth numbers of periods. It is a scalar: zero indicates the correct count, while larger values indicate a greater mismatch without distinguishing overprediction from underprediction.

Period correspondence. ARI and AMI evaluate the agreement of predicted and ground-truth periods as a whole. However, ranges, period descriptions, and period order require correspondence measures between individual predicted and ground-truth periods. Period indices cannot be paired directly when periods are missing, added, split, or merged. We therefore employ Hungarian matching (Kuhn, 1955) based on the overlap between the images assigned to each pair of periods. It selects the set of one-to-one matches that maximises total image overlap across all selected pairs. Matches are therefore chosen jointly rather than by assigning each prediction independently to its highest-overlap ground-truth period. Periods sharing no images remain unmatched. The same pairing is used for all period-level comparisons. Hungarian matching is an alignment step rather than a metric, so it is not listed in Tab. 4. Details can be found in App. E.1.

Period ranges and text. Each predicted range spans the earliest and latest supplied years among its assigned images: t_i in independent evaluation and $\hat{t}_i$ in end-to-end evaluation. For each matched pair, temporal intersection over union (temporal IoU) compares the predicted and ground-truth ranges, while BERTScore-F1 and MoverScore compare the period titles and descriptions.

Trajectory order and completeness. At the topic level, title and description are compared directly with their single references using the text metrics. At the trajectory level, Kendall's τ (Kendall, 1938) compares the order of matched periods and requires at least two matched pairs; we report its macro-average over eligible topics and their count. For the complete trajectory, matched predictions are encoded by their ground-truth period indices and unmatched predictions by distinct placeholders. Normalised Levenshtein distance compares this encoding with the ground-truth sequence, capturing missing, additional, and misordered periods. Sections E.2 and E.3 provide both definitions.

Table 4: Metrics and baseline results on VDBench. Task 2 scores are macro-averaged over topics.

Task 1: Visual context detection

	Output	Metric	OPEN-MLLM	TRAJEX
$\hat{t}_i$	Year	MAE (years) ↓	4.72	**4.65**
$\hat{e}_i$	Environment	Accuracy / macro-F1 ↑	0.26 / 0.21	**0.27 / 0.26**
$\hat{a}_i$	Activity	Accuracy / macro-F1 ↑	0.28 / 0.24	**0.31 / 0.32**
$\hat{c}_i$	Caption	BERTScore / MoverScore ↑	0.84 / 0.12	**0.86 / 0.14**

Task 2: Topic and trajectory discovery

			OPEN-MLLM		TRAJEX	
	Output	**Metric**	**GT**	**Pred.**	**GT**	**Pred.**
Trajectory structure						
$\hat{\mathcal{I}}_k$	Image grouping	ARI / AMI ↑	0.56 / 0.66	0.41 / 0.54	**0.79 / 0.87**	**0.46 / 0.58**
$\hat{J}$	Period count	$\lvert\hat{J} - J\rvert$ ↓	**0.63**	**0.93**	1.17	2.02
$\hat{R}_k$	Period range	Temporal IoU ↑	0.46	0.35	**0.63**	**0.39**
$\hat{\mathcal{P}}$	Matched order	Kendall's τ ↑	0.97	0.75	**1.00**	**1.00**
$\hat{\mathcal{P}}$	Full sequence	Normalised Levenshtein ↓	0.24	0.45	**0.20**	**0.41**
Trajectory semantics						
$\hat{s}_k$	Period title	BERTScore / MoverScore ↑	**0.85 / 0.09**	**0.84 / 0.02**	0.84 / 0.03	0.84 / 0.01
$\hat{m}_k$	Period description	BERTScore / MoverScore ↑	**0.86 / 0.14**	**0.85 / 0.11**	0.85 / 0.12	0.84 / 0.10
$\hat{n}$	Topic title	BERTScore / MoverScore ↑	**0.89 / 0.35**	**0.89 / 0.34**	0.87 / 0.25	0.87 / 0.23
$\hat{d}$	Topic description	BERTScore / MoverScore ↑	**0.84 / 0.05**	0.83 / 0.03	**0.84 / 0.05**	**0.84 / 0.04**

6 BASELINES AND RESULTS

We evaluate two zero-shot, open-weight reference baselines: OPEN-MLLM, a single-model approach, and TRAJEX (*Trajectory Explorer*), a multi-component pipeline. Neither uses VDBench for training, fine-tuning, or prompt selection. In Task 2, each receives either ground truth (GT) Task 1 values or its own Task 1 predictions (Pred.). Implementation details are provided in App. F.

OPEN-MLLM. Open-MLLM applies Qwen3-VL-32B-Instruct (Qwen Team, 2025) directly to both tasks, without auxiliary encoders or clustering. For Task 1, it predicts the year, environment, activity, and caption of each image independently. For Task 2, it uses the supplied Task 1 values instead of processing the images again. The model predicts the topic title and description, determines the periods, assigns the images to them, and generates their titles and descriptions. Ablations in App. H.6, H.7, and H.11 cover model size, image perturbations, each Task 1 input, and the true period count.

TRAJEX. TrajEx instead assigns outputs to specialised components. For Task 1, the method of Tekaya et al. (2025) estimates the year with a frozen EVA-CLIP encoder (Sun et al., 2023). EVA-CLIP zero-shot similarity predicts the environment and activity, while Qwen3-VL-8B-Instruct generates the caption. For Task 2, the images are ordered by their supplied GT or predicted years and divided into consecutive periods using dynamic-programming change-point detection (Truong et al., 2020). It also sets the number of periods. DINOv3 features (Siméoni et al., 2025) and the supplied Task 1 values are then used to merge small periods, while the visual features also select representative images. Qwen3-VL-8B-Instruct uses the resulting periods to generate the period and topic texts. The appendix gives the complete implementation, tests its dependence on the Task 1 inputs, and compares its year-based division with feature-space clustering (App. F, H.7, and H.8).

Results. Table 4 reports both tasks. Task 1 scores are close. TrajEx is numerically better on every output, but only its activity-accuracy advantage is consistent across topics. With GT inputs, TrajEx performs better on image grouping and period ranges, while Open-MLLM achieves higher period-text MoverScore. With predicted inputs, only period-count error and period-title MoverScore differ consistently, both in favour of Open-MLLM. Moving from GT to predicted inputs lowers ARI by 0.337 for TrajEx but only by 0.153 for Open-MLLM. The ablations explain this gap: TrajEx depends on accurate years and period-count selection, and its year-based division is essential, as feature-space clustering reduces its GT-input ARI from 0.794 to 0.129. Open-MLLM relies more on captions and remains limited by image assignment, with only a 0.013 ARI gain from the correct period count. The aggregate scores also hide topic-level differences: with predicted inputs, Open-MLLM has the higher ARI on 19 of the 41 topics. Full results and ablations are provided in App. G and H.

7 Limitations and Future Directions

This work represents a first step towards formalising VDA and enabling a quantitative evaluation of VDA methods. It is limited in the following respects, which opens up future research directions. *First*, VDBench uses synthetic rather than real archival images. This choice was made to make a first systematic evaluation of VDA methods possible. This would not work yet with real archive images due to their sparsity and missing metadata. Future research will focus on adding a validation dataset with real images to VDBench to enable an independent evaluation of VDA methods on real material. *Second*, expert reviews are limited. While all included trajectories were expert-reviewed, only a random sample of images could be verified for plausibility due to time constraints. To compensate for this limitation, we applied automated consistency and verification checks for all images (Sec. 3). *Third*, there are sources of potential bias in the collection. The LLMs employed for topic construction introduced a Western bias into the collection, i.e. Western values, histories, and interpretations. Furthermore, potential biases may arise from web search and zero-shot CLIP-based filtering. In future work, our goal is to quantify these biases and develop diversification strategies. *Fourth*, BERTScore and MoverScore measure agreement with reference texts which have not been specifically fine-tuned to the historical domain. Future work may focus on dedicated measures to more accurately measure semantic similarity in historical contexts. Still, the employed measures enable a fair comparison, which has been the priority for the benchmark.

8 Related Work

Visual diachronic analysis borrows its question from lexical semantic change detection, which traces how the meaning of a word shifts by comparing its uses across dated text collections (Kutuzov et al., 2018; Tahmasebi et al., 2021). SemEval-2020 Task 1 gave this line of work a shared evaluation (Schlechtweg et al., 2020), and later SemEval tasks brought images in by linking word senses and idioms to them (Raganato et al., 2023; Pickard et al., 2025), yet without any notion of time. Computer vision has mostly treated time as something to predict or to be robust to. Images are dated from their appearance (Ginosar et al., 2017; Salem et al., 2016), visual change is mined across large collections (Matzen et al., 2017; Deng et al., 2025), and vision-language models encode the year of an image (Tekaya et al., 2025) but degrade as their training data ages (Garg et al., 2024). What none of these works does is describe how the visual context of one subject evolves from period to period.

Evaluating this requires images of the same subject for every documented period, which archives rarely provide. Synthetic data is an established answer to such scarcity, and generated images can replace real training data in image recognition (He et al., 2023) and human body estimation (Black et al., 2023). For the past, however, text-to-image models produce anachronisms and era-specific default styles (De Rosa Palmini & Cetinic, 2026), so generated images are only usable as benchmark data if grounded and validated. To the best of our knowledge, no existing benchmark combines source-grounded historical periods with per-image context labels to evaluate both individual depictions and complete diachronic trajectories. App. I covers the related work in full.

9 Conclusion

We introduced VDBench, the first benchmark for analysing *visual diachrony*, i.e. how the visual context of a recognisable subject changes across historical periods. VDBench evaluates both the context depicted in individual images and a model's ability to discover how that context evolves over time. Expert validation supports the historical relevance of the compiled trajectories and the visual plausibility of the generated images. First baselines built on top of state-of-the-art foundation models underline the complexity of the task, making it an attractive challenge for future research.

Future work. Future releases should broaden the geographic and cultural coverage, obtain multiple independent reviews per topic, and incorporate licensed archival images with reliable annotations. Methodological work should study hybrid evaluation on synthetic and archival data, uncertainty-aware trajectory modelling, and text evaluation supported by multiple references and human assessment.

REFERENCES

Waleed Ammar, Dirk Groeneveld, Chandra Bhagavatula, Iz Beltagy, Miles Crawford, Doug Downey, Jason Dunkelberger, Ahmed Elgohary, Sergey Feldman, Vu Ha, Rodney Kinney, Sebastian Kohlmeier, Kyle Lo, Tyler Murray, Hsu-Han Ooi, Matthew Peters, Joanna Power, Sam Skjonsberg, Lucy Lu Wang, Chris Wilhelm, Zheng Yuan, Madeleine van Zuylen, and Oren Etzioni. Construction of the literature graph in semantic scholar. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 3 (Industry Papers)*, pp. 84–91. Association for Computational Linguistics, 2018. doi: 10.18653/v1/N18-3011. 16

Diane Atkinson. *Rise Up, Women!: The Remarkable Lives of the Suffragettes*. Bloomsbury Publishing, 2018. 2

Alexandra Barancová, Melvin Wevers, and Nanne van Noord. Blind dates: Examining the expression of temporality in historical photographs. In *Proceedings of the Computational Humanities Research Conference (CHR)*, volume 3558 of *CEUR Workshop Proceedings*, pp. 490–499. CEUR-WS, 2023. arXiv:2310.06633. 2, 44

Mary Bazire and Patrick Brézillon. Understanding context before using it. In *International and Interdisciplinary Conference on Modeling and Using Context*, pp. 29–40. Springer, 2005. 3

Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. Easily accessible text-to-image generation amplifies demographic stereotypes at large scale. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pp. 1493–1504. ACM, 2023. doi: 10.1145/3593013.3594095. 44

Michael J. Black, Priyanka Patel, Joachim Tesch, and Jinlong Yang. BEDLAM: A synthetic dataset of bodies exhibiting detailed lifelike animated motion. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8726–8737, 2023. 9, 44

Andreas Blank. Why do new meanings occur? a cognitive typology of the motivations for lexical semantic change. In Andreas Blank and Peter Koch (eds.), *Historical Semantics and Cognition*, pp. 61–90. Mouton de Gruyter, Berlin, 1999. doi: 10.1515/9783110804195.61. 43

Patrick Brézillon. Context in problem solving: A survey. *The Knowledge Engineering Review*, 14(1):47–80, 1999. 3

Ricardo J. G. B. Campello, Davoud Moulavi, and Joerg Sander. Density-based clustering based on hierarchical density estimates. In *Advances in Knowledge Discovery and Data Mining*, volume 7819 of *Lecture Notes in Computer Science*, pp. 160–172. Springer, 2013. doi: 10.1007/978-3-642-37456-2_14. 40

Centre for Time Use Research. Multinational time use study (MTUS) user guide. Release 7.0 of the World 5.53 dataset, https://www.timeuse.org/MTUS-User-Guide, 2020. October 2020. 4, 18

Eric Ming Chen, Jin Sun, Apoorv Khandelwal, Dani Lischinski, Noah Snavely, and Hadar Averbuch-Elor. What's in a decade? transforming faces through time. *Computer Graphics Forum*, 42(2):281–291, 2023. doi: 10.1111/cgf.14761. 44

Patricia A Cunningham. *Reforming Women's Fashion, 1850–1920: Politics, Health, and Art*. Kent State University Press, 2003. 2

Maria-Teresa De Rosa Palmini and Eva Cetinic. Synthetic history: Evaluating visual representations of the past in diffusion models. In *International Conference on Learning Representations (ICLR)*, 2026. arXiv:2505.17064. 9, 44

DeepSeek-AI. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3, 16

Boyang Deng, Songyou Peng, Kyle Genova, Gordon Wetzstein, Noah Snavely, Leonidas Guibas, and Thomas Funkhouser. Visual chronicles: Using multimodal LLMs to analyze massive collections of images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 12769–12778, 2025. doi: 10.1109/ICCV51701.2025.01186. 2, 9, 43

Anind K Dey, Gregory D Abowd, and Daniel Salber. A conceptual framework and a toolkit for supporting the rapid prototyping of context-aware applications. *Human–Computer Interaction*, 16(2-4):97–166, 2001. 3

Shiv Ram Dubey. A decade survey of content based image retrieval using deep learning. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(5):2687–2704, 2022. doi: 10.1109/TCSVT.2021.3080920. 3

DuckDuckGo. Duckduckgo image search. https://duckduckgo.com/, 2026. Accessed September 2026. 4, 18

Eurostat. Harmonised european time use surveys (HETUS) 2018 guidelines, 2020. 2020 edition, KS-GQ-20-011. 4, 18

Carolina Galleguillos and Serge Belongie. Context-based object categorization: A critical survey. *Computer Vision and Image Understanding*, 114(6):712–722, 2010. 2, 43

Saurabh Garg, Mehrdad Farajtabar, Hadi Pouransari, Raviteja Vemulapalli, Sachin Mehta, Oncel Tuzel, Vaishaal Shankar, and Fartash Faghri. TiC-CLIP: Continual training of CLIP models. In *International Conference on Learning Representations (ICLR)*, 2024. 9, 44

Shiry Ginosar, Kate Rakelly, Sarah M. Sachs, Brian Yin, Crystal Lee, Philipp Krähenbühl, and Alexei A. Efros. A century of portraits: A visual historical record of american high school yearbooks. *IEEE Transactions on Computational Imaging*, 3(3):421–431, 2017. doi: 10.1109/TCI.2017.2699865. 9, 43

Google DeepMind. Introducing Nano Banana Pro. https://blog.google/innovation-and-ai/products/nano-banana-pro/, 2025. Accessed: 2026-09-17. 4, 18

William L Hamilton, Jure Leskovec, and Dan Jurafsky. Diachronic word embeddings reveal statistical laws of semantic change. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1489–1501. Association for Computational Linguistics, 2016. 1, 43

Ruifei He, Shuyang Sun, Xin Yu, Chuhui Xue, Wenqing Zhang, Philip Torr, Song Bai, and Xiaojuan Qi. Is synthetic data from generative models ready for image recognition? In *International Conference on Learning Representations (ICLR)*, 2023. 9, 44

Ginny Hendricks, Dominika Tkaczyk, Jennifer Lin, and Patricia Feeney. Crossref: The sustainable source of community-owned scholarly metadata. *Quantitative Science Studies*, 1(1):414–427, 2020. doi: 10.1162/qss_a_00022. 16

Lawrence Hubert and Phipps Arabie. Comparing partitions. *Journal of Classification*, 2(1):193–218, 1985. 7

Ramesh Jain and Pinaki Sinha. Content without context is meaningless. In *Proceedings of the 18th ACM International Conference on Multimedia*, pp. 1259–1268. ACM, 2010. 3

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. Beyond aesthetics: Cultural competence in text-to-image models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2024. doi: 10.52202/079017-0439. 44

Maurice G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1–2):81–93, 1938. doi: 10.1093/biomet/30.1-2.81. 7

Minchul Kim, Feng Liu, Anil Jain, and Xiaoming Liu. DCFace: Synthetic face generation with dual condition diffusion model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 12715–12725, 2023. 44

Harold W. Kuhn. The Hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2 (1–2):83–97, 1955. 7, 28

Andrey Kutuzov, Lilja Øvrelid, Terrence Szymanski, and Erik Velldal. Diachronic word embeddings and semantic shifts: A survey. In *Proceedings of the 27th International Conference on Computational Linguistics*, pp. 1384–1397, Santa Fe, New Mexico, USA, 2018. Association for Computational Linguistics. 9, 43

Yong Jae Lee, Alexei A. Efros, and Martial Hebert. Style-aware mid-level representation for discovering visual connections in space and time. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, pp. 1857–1864. IEEE, 2013. doi: 10.1109/ICCV.2013.233. 43

Hongsheng Li, Guangming Zhu, Liang Zhang, Youliang Jiang, Yixuan Dang, Haoran Hou, Peiyi Shen, Xia Zhao, Syed Afaq Ali Shah, and Mohammed Bennamoun. Scene graph generation: A comprehensive survey. *Neurocomputing*, 566:127052, 2024. 43

Weixin Liang and James Zou. Metashift: A dataset of datasets for evaluating contextual distribution shifts and training conflicts. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2202.06523. 2, 43

Zhiqiu Lin, Jia Shi, Deepak Pathak, and Deva Ramanan. The CLEAR benchmark: Continual LEArning on real-world imagery. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, 2021. arXiv:2201.06289. 44

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 36, 2023. arXiv:2303.11408. 44

Xuan Luo, Xuaner Zhang, Paul Yoo, Ricardo Martin-Brualla, Jason Lawrence, and Steven M. Seitz. Time-travel rephotography. *ACM Transactions on Graphics*, 40(6):213:1–213:12, 2021. doi: 10.1145/3478513.3480485. 44

Kevin Matzen, Kavita Bala, and Noah Snavely. StreetStyle: Exploring world-wide clothing styles from millions of photos, 2017. arXiv:1706.01869. 9, 43

Open Library. Open library apis. https://openlibrary.org/developers/api, 2026. Accessed September 2026. 16

OpenAI. GPT-5.6 Sol Model. https://developers.openai.com/api/docs/models/gpt-5.6-sol, 2026. Accessed September 2026. 4, 18

Frank Palermo, James Hays, and Alexei A. Efros. Dating historical color images. In *Computer Vision – ECCV 2012*, volume 7577 of *Lecture Notes in Computer Science*, pp. 499–512. Springer, 2012. doi: 10.1007/978-3-642-33783-3_36. 43

Thomas Pickard, Aline Villavicencio, Maggie Mi, Wei He, Dylan Phelps, and Marco Idiart. SemEval-2025 task 1: AdMIRe – advancing multimodal idiomaticity representation. In *Proceedings of the 19th International Workshop on Semantic Evaluation (SemEval-2025)*, pp. 2597–2609, Vienna, Austria, July 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.semeval-1.330/. 9, 43

Jason Priem, Heather Piwowar, and Richard Orr. OpenAlex: A fully-open index of scholarly works, authors, venues, institutions, and concepts. *arXiv preprint arXiv:2205.01833*, 2022. doi: 10.48550/arXiv.2205.01833. 16

Qwen Team. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 8, 29

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pp. 8748–8763, 2021. 4, 18

Alessandro Raganato, Iacer Calixto, Asahi Ushio, Jose Camacho-Collados, and Mohammad Taher Pilehvar. SemEval-2023 task 1: Visual word sense disambiguation. In *Proceedings of the 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pp. 2227–2234, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.semeval-1.308. 9, 43

Tawfiq Salem, Scott Workman, Menghua Zhai, and Nathan Jacobs. Analyzing human appearance as a cue for dating images. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 1–8. IEEE, 2016. 9, 43

Ferdinand de Saussure. *Course in General Linguistics*. Philosophical Library, New York, 1959. Edited by Charles Bally and Albert Sechehaye in collaboration with Albert Riedlinger; translated by Wade Baskin. 1, 43

Grant Schindler, Frank Dellaert, and Sing Bing Kang. Inferring temporal order of images from 3D structure. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2007. 43

Dominik Schlechtweg, Barbara McGillivray, Simon Hengchen, Haim Dubossarsky, and Nina Tahmasebi. SemEval-2020 task 1: Unsupervised lexical semantic change detection. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pp. 1–23. International Committee for Computational Linguistics, 2020. doi: 10.18653/v1/2020.semeval-1.1. 9, 43

David Semedo and João Magalhães. Diachronic cross-modal embeddings. In *Proceedings of the 27th ACM International Conference on Multimedia*, pp. 2061–2069. ACM, 2019. 44

Jamie Shotton, Andrew Fitzgibbon, Mat Cook, Toby Sharp, Mark Finocchio, Richard Moore, Alex Kipman, and Andrew Blake. Real-time human pose recognition in parts from single depth images. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1297–1304. IEEE, 2011. doi: 10.1109/CVPR.2011.5995316. 44

Oriane Siméoni et al. DINOv3. *arXiv preprint arXiv:2508.10104*, 2025. 8, 30

Jennifer Stuber, Sandro Galea, and Bruce G. Link. Smoking and the emergence of a stigmatized social status. *Social Science & Medicine*, 67(3):420–430, 2008. doi: 10.1016/j.socscimed.2008.03.010. 3

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. EVA-CLIP: Improved training techniques for CLIP at scale. *arXiv preprint arXiv:2303.15389*, 2023. 8, 30

Nina Tahmasebi, Lars Borin, and Adam Jatowt. Survey of computational approaches to lexical semantic change detection. In Nina Tahmasebi, Lars Borin, Adam Jatowt, Yang Xu, and Simon Hengchen (eds.), *Computational Approaches to Semantic Change*, pp. 1–91. Language Science Press, Berlin, 2021. doi: 10.5281/zenodo.5040302. 1, 9, 43

Nidham Tekaya, Manuela Waldner, and Matthias Zeppelzauer. A matter of time: Revealing the structure of time in vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 12371–12380. ACM, 2025. doi: 10.1145/3746027.3758163. 2, 8, 9, 19, 30, 44

Elizabeth Closs Traugott and Richard B. Dasher. *Regularity in Semantic Change*. Cambridge Studies in Linguistics. Cambridge University Press, Cambridge, 2002. ISBN 0521583780. 1, 43

Charles Truong, Laurent Oudre, and Nicolas Vayatis. Selective review of offline change point detection methods. *Signal Processing*, 167:107299, 2020. 8, 30

Nan Turner. *Clothing Goes to War: Creativity Inspired by Scarcity in World War II*. Intellect Books, 2021. 2

U.S. Bureau of Labor Statistics. American time use survey (ATUS) activity coding lexicons. https://www.bls.gov/tus/lexicons.htm, 2023. Accessed September 2026. 4, 18

Nguyen Xuan Vinh, Julien Epps, and James Bailey. Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. *Journal of Machine Learning Research*, 11: 2837–2854, 2010. 7

Sirion Vittayakorn, Alexander C. Berg, and Tamara L. Berg. When was that made? In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, pp. 715–724. IEEE, 2017. doi: 10.1109/WACV.2017.85. 43

Ziyu Wan, Bo Zhang, Dongdong Chen, Pan Zhang, Dong Chen, Jing Liao, and Fang Wen. Bringing old photos back to life. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2747–2757, 2020. doi: 10.1109/CVPR42600.2020.00282. 44

Xuan Wang and Zhigang Zhu. Context understanding in computer vision: A survey. *Computer Vision and Image Understanding*, 229:103646, 2023. 2, 43

Wikipedia contributors. Wikipedia category trees for photographic equipment, film formats, and dated cultural phenomena (cameras and camera lenses by year of introduction, film format, timeline of photography technology, and related fads and trends categories). Wikipedia category trees, https://en.wikipedia.org/wiki/Category:Cameras_by_year_of_introduction, 2026. Accessed September 2026. 4, 18

Jianxiong Xiao, James Hays, Krista A. Ehinger, Aude Oliva, and Antonio Torralba. SUN database: Large-scale scene recognition from abbey to zoo. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3485–3492. IEEE, 2010. 4, 18

Huaxiu Yao, Caroline Choi, Bochuan Cao, Yoonho Lee, Pang Wei Koh, and Chelsea Finn. Wild-Time: A benchmark of in-the-wild distribution shift over time. In *Advances in Neural Information Processing Systems (NeurIPS), Datasets and Benchmarks Track*, volume 35, 2022. arXiv:2211.14238. 44

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 5534–5542. IEEE, 2016. 43

Mert Yuksekgonul, Federico Bianchi, Pratyusha Kalluri, Dan Jurafsky, and James Zou. When and why vision-language models behave like bags-of-words, and what to do about it? In *International Conference on Learning Representations*, 2023. URL https://arxiv.org/abs/2210.01936. 2, 43

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. BERTScore: Evaluating text generation with BERT. In *International Conference on Learning Representations (ICLR)*, 2020. 7

Wei Zhao, Maxime Peyrard, Fei Liu, Yang Gao, Christian M. Meyer, and Steffen Eger. MoverScore: Text generation evaluating with contextualized embeddings and earth mover distance. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 563–578, 2019. doi: 10.18653/v1/D19-1053. 7

Bolei Zhou, Agata Lapedriza, Aditya Khosla, Aude Oliva, and Antonio Torralba. Places: A 10 million image database for scene recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 40(6): 1452–1464, 2018. doi: 10.1109/TPAMI.2017.2723009. 4, 18, 43

APPENDIX OVERVIEW

App.	Supports	Content
A	Dataset (Sec. 3)	Construction funnel, Historian and Librarian roles, source audit, record schema, released splits
B	Dataset (Sec. 3)	Vocabularies, zero-shot checks, reference conditioning, retries
C	Dataset (Sec. 3)	The four pipeline prompts and one compiled generation prompt
D	Validation (Sec. 4)	Aggregation rule, Stage 1 and Stage 2 records, interfaces, scope of the evidence
E	Evaluation (Sec. 5.2)	Period alignment, temporal IoU, sequence distance, worked example
F	Baselines (Sec. 6)	Component map and implementation of Open-MLLM and TrajEx
G	Results (Sec. 6)	Per-topic scores, assignment patterns, matched-period examples
H	Results (Sec. 6)	Ten ablation and diagnostic studies
I	Related work (Sec. 8)	Extended discussion and the boundary between VDA and adjacent tasks

Three conventions hold throughout. Topic names are printed in italics and are the titles stored in the dataset; the topic the main paper calls *women in trousers* is stored as *Women wearing trousers*. GT denotes the setting in which Task 2 receives ground-truth Task 1 values and Pred. the setting in which it receives a baseline's own Task 1 predictions (Sec. 5.2). The symbol "–" in a table marks a quantity that is not reported, never a value of zero, and every caption states why.

A DATASET CONSTRUCTION DETAILS

A.1 FROM PROPOSED TOPICS TO GENERATED IMAGES

Table 5 gives the count entering and leaving every stage, so the 20,100 released images can be traced back to the 266 topics originally proposed.

Table 5: Construction funnel. Each row states what the stage decides and how many topics survive it.

Stage	Decision	Outcome
1. Topic proposal	*Historian* proposes photographable topics, their periods, visible characteristics, and historical meanings	266 topics proposed
2. Source retrieval and linking	*Librarian* retrieves candidate sources and links each to the claim it supports	Sources attached to period claims
3. Author source audit	Authors verify bibliographic records and remove topics whose claims are unsupported or whose sources cannot be identified	148 topics retained (55.6%), 118 removed; 706 period descriptions; 174 verified sources
4. Stage 1 expert review	Experts select topics within their expertise and judge visible change, cultural relevance, period boundaries, and period descriptions	46 of the 148 topics reviewed; 41 approved, 3 rejected, 2 tied
5. Controlled image generation	100 images generated and verified for every period of an approved topic	41 topics, 201 periods, 20,100 images

Expert review is not exhaustive over the 148 audited topics: reviewers browsed the full list and selected only topics inside their own field, so 46 were reviewed and the remaining 102 were never generated. The 41 approved topics carry four to six periods each, averaging $201/41 = 4.90$, the ground-truth mean used in App. H.8.

A.2 Historian and Librarian Roles

DeepSeek-R1 (DeepSeek-AI, 2025) runs in two separate roles with no shared context. As *Historian* it receives the 1840 photographic year floor and the titles already proposed, and returns photographable subjects whose documented context or meaning changes while the subject stays recognisable. Each candidate carries an ordered sequence of periods with year ranges, short context labels, historical meanings, and claim-specific search terms. The prompt forbids inventing publications, DOIs, quotations, or URLs, and instructs the model to omit any change it cannot state without speculation (Listing 1).

As *Librarian* it receives one candidate together with the documents retrieved for its claims. It links a document only to the period or transition that the document supports, rejects generic search-result URLs and mismatched works, marks unsupported claims explicitly, and recommends retaining, revising, or rejecting the candidate. A source is admissible only with a locatable DOI or direct URL. The role does not assign verified status; that is decided by the audit below (Listing 2).

A.3 Manual Source Audit

Both agent outputs are proposals, not accepted evidence. The authors check manually whether the cited material supports each period and historical claim rather than merely mentioning the subject. Bibliographic records are resolved primarily through Crossref (Hendricks et al., 2020), with OpenAlex (Priem et al., 2022), Semantic Scholar (Ammar et al., 2018), Open Library (Open Library, 2026), publisher pages, and repository records used when a record cannot be resolved there. Most retained sources are scholarly articles and books; reference, institutional, archival, and general webpages, including Wikipedia, provide supplementary evidence. A topic is removed when any of its periods rests on a claim that no retained source supports, or when a cited work cannot be identified at all. This audit is what reduces 266 proposed topics to 148.

A.4 Record Structure

Table 6: Principal fields in a VDBench record. Construction metadata document how an image was produced; they are not prediction targets and are not supplied to any method.

Record level	**Fields**
Topic	Identifier, title, overall description, ordered periods, and verified sources
Period	Title, inclusive year range, visible characteristics, and documented meaning
Image	Identifier, parent topic and period, depicted year, environment, activity, and caption
Construction metadata	Image prompt, model identifier, dated trend if any, selected reference images, retry history, and verification results

The hierarchy in Tab. 6 is strict: each image belongs to exactly one period and each period to exactly one topic, which is what makes the ground-truth periods a partition of a topic's images (Sec. 5.1). The caption field carries one property that bounds what Task 1 caption scores can mean. It is written together with the generation prompt, so it records the *intended* content of the image rather than an independent description of the rendered result: a method that reproduces the caption is reproducing the specification the renderer was given, not a human annotation. App. H.10 measures one consequence, that a quarter of the captions state a year or decade explicitly.

A.5 Released Splits

Open-vocabulary methods, including both baselines in this paper, use the complete collection and need no VDBench training; the splits exist for supervised submissions and are used nowhere in the results reported here. The release defines two schemes over the same 20,100 images, per image, in `splits/vdbench_splits.json` and `splits/vdbench_records.csv`. Table 7 gives their sizes. Scores on the two are not comparable, so a supervised submission must state which one it used.

Topic-disjoint partitions whole topics, so every image of a held-out topic is unseen and the scheme measures transfer to diachronic topics never observed. The allocation is an exact count rather than a

Table 7: The two released split schemes over the same 20,100 images. Shares are percentages of the collection. For configuration-disjoint, periods are 201 less the topic-periods with too few configuration groups to divide, which the release records explicitly rather than dropping.

Scheme	Partition	Topics	Periods	Images	Share
Topic-disjoint	Train	29	140	14,000	69.7%
(`seed 42`)	Val	6	30	3,000	14.9%
	Test	6	31	3,100	15.4%
Configuration-disjoint	Train	41	201	14,326	71.3%
(`seed 43`)	Val	41	200	2,167	10.8%
	Test	41	197	3,607	17.9%

ratio, 29/6/6 topics, assigned by a deficit-based round robin over topics sorted by period count and then year span. The validation topics are *Blue jeans / denim trousers*, *Female body weight ideals*, *Billboard / outdoor advertising*, *Kitchen design and layout*, *Automobile design*, and *Pink and blue gender coding*; the test topics are *Women wearing trousers*, *Corset and waist-cinching*, *Cigarette / smoking*, *Selfie / self-portrait photograph*, *Bicycle*, and *Buddha statue in Western home*. The main paper's running example and two of the four gallery topics of App. G.2 are therefore test topics, which involves no leakage here because both baselines are zero-shot on the complete collection.

One property of this draw should be checked before the scheme is used. Balancing succeeded on period count, 4.83 periods per training topic against 5.08 per held-out topic, but not on year span: training topics span 106.3 years on average against 167.1 for the twelve held-out ones, every held-out topic spans at least 139 years, and 23 of the 29 training topics span less than that. The round robin distributes the widest-spanning topics first across all three partitions and then fills the remaining training slots from the short-span tail. Held-out topics consequently have longer periods lying further apart in time, which is the easier regime for any method that segments on an estimated year (App. H.7), so topic-disjoint scores should not be read against scores on the complete collection.

Configuration-disjoint is the scheme Sec. 3 refers to as the uniform split: all 41 topics appear in all three partitions, and it evaluates generalisation to new visual configurations of topics already seen. It is stricter than a uniform draw over images. Images are grouped by topic, period, environment, activity, and dated trend, the generation parameters of App. B.2, and each group is assigned whole, so near-duplicate images produced from one configuration cannot straddle training and test. Groups are distributed per topic and per period rather than pooled over a topic's whole run, so period coverage survives wherever a period has enough distinct groups. Two consequences are worth stating. Because group sizes vary, the realised sizes depart from the 70/10/20 target, most visibly in the test partition at 17.9%. And four topic-periods had too few groups to divide: the 1960s–1970s period of *Short women's hair (bob cut)* is absent from validation and test, and three periods of *Meme / viral image* are absent from test. The release records these in `absent_topic_periods` rather than letting them land silently in one partition, but a trajectory-level metric computed on the test partition of this scheme is computed over an incomplete trajectory for those two topics.

B Image-Generation Details

B.1 Visual Vocabularies

Environment and activity are the controlled context dimensions and are released as per-image labels, so they are also the Task 1 label spaces (Sec. 5.1). The dated trend is an optional time-specific cue that enters 5% of the generation prompts and is kept only as construction metadata; no task asks a method to predict it.

B.2 Specifying One Image

The topic and period are fixed beforehand; everything below varies across the 100 images of a period, which is what makes the collection diverse while keeping the diachronic signal controlled.

1. **Draw a combination.** One environment from the 669 scene classes and one activity from the 177 time-use classes; with probability 5% a dated trend as well.

Table 8: Sizes, sources, and examples of the three vocabularies used to vary visual context.

Vocabulary	Size	Compiled from	Examples
Environment	669	Places365 (Zhou et al., 2018), SUN (Xiao et al., 2010)	kitchen, classroom, assembly line, street, gas station, forest path, beach
Activity	177	ATUS (U.S. Bureau of Labor Statistics, 2023), HETUS (Eurostat, 2020), MTUS (Centre for Time Use Research, 2020)	cooking, caring for children, working, shopping, socialising, reading, commuting
Dated trend	74	dated Wikipedia categories (Wikipedia contributors, 2026)	Charleston (1923), hula hoop (1958), breakdancing (1983), flash mob (2003), Ice Bucket Challenge (2014)

2. **Ground it in photographs.** DuckDuckGo Image Search (DuckDuckGo, 2026) retrieves candidate photographs for the topic, environment, and activity, and CLIP (Radford et al., 2021) tests whether they are present (App. B.3). Combinations without photographic support are discarded before anything is rendered.
3. **Write the specification.** GPT-5.6 Sol (`gpt-5.6-sol`) (OpenAI, 2026) receives the reviewed period description with its start and end years and the drawn combination, picks one year inside the period, and writes an image-generation prompt plus a caption for the intended content (Listing 3).
4. **Render.** Nano Banana Fast (Google DeepMind, 2025) renders the image, for half of the specifications conditioned on up to two retrieved reference photographs (App. B.4).
5. **Verify and retry.** CLIP re-applies the checks to the rendered image. An image whose selected elements are not detected is regenerated, and after repeated failure the pipeline moves to another combination that passed step 2. Only images passing all assigned checks are retained.

Table 9: Fixed image-generation settings.

Setting	Value
Images per period	100
Prompts containing a dated trend	5%
Image specifications for which reference conditioning is attempted	50%
Reference images supplied to one generation	At most 2
Reference-selection attempts per specification	At most 8

B.3 Zero-Shot Visual Checks

CLIP is used as a *ranker*, not as a thresholded detector, which avoids a similarity cut-off that would need per-topic calibration. For an environment check the candidate set is the topic name together with all 669 environment labels; CLIP scores every candidate against the image, and the check passes only if the topic name and the selected environment occupy the top two positions, in either order. The activity check is identical with the 177 activity labels, and the dated-trend check with the 74 trend labels.

The same probe runs at two points with different inputs and different consequences. Before generation it filters plausibility on real photographs: a topic–environment–activity combination proceeds only if retrieved photographs show it, which removes combinations that would force the renderer to invent an implausible scene, and a failure discards the combination unrendered. After generation it verifies the rendered image, and a failure triggers regeneration and, after repeated failure, another grounded combination. The optional dated trend is excluded from the pre-generation filter, because requiring one retrieved photograph to contain the trend as well would eliminate most combinations; it is checked only after rendering.

B.4 Reference-Conditioned Generation

Half of the specifications attempt reference-conditioned generation, supplying Nano Banana Fast with up to two retrieved photographs alongside the prompt. CLIP screens each candidate reference for the depicted subject and for the year the model selected, with at most eight attempts before falling back to generation without a reference. Using CLIP for the year is motivated by the finding that CLIP representations encode temporal structure (Tekaya et al., 2025). This year check governs *reference selection only* and is never applied to the generated image, so no generated image is filtered on its own apparent date and the released year label remains the one drawn from the documented period.

C Pipeline Prompts

The prompts record which content was model-generated before the source audit and the image checks, and therefore which parts of VDBench are machine-proposed rather than documented. Listings 1 and 2 are the two DeepSeek-R1 roles of stages 1 and 2 in Tab. 5; Listing 3 is the GPT-5.6 Sol specification template of stage 5, whose {{name}} placeholders are filled from the reviewed topic record and the drawn generation parameters; and Listing 4 is one such prompt after substitution, as supplied to Nano Banana Fast.

Listing 1: Historian proposal prompt.

```
SYSTEM ROLE
You are the Historian agent for the VDBench visual diachrony dataset. Propose
photographable visual elements whose cultural meaning or social reading
changes across time. A topic is valid only when the same recognizable
element persists while its documented context or meaning changes.

INPUT
- photographic year floor: 1840
- existing topic titles: {{existing_titles}}
- requested number of candidates: {{n_candidates}}

TASK
1. Propose non-duplicate visual topics.
2. Divide each topic into a concise sequence of historically defensible,
   era-defining periods.
3. For every period give a year range, short context label, and concise
   account of what the element meant or signified in that interval.
4. Emit claim-specific search terms for the topic and each transition.
5. Do not invent publications, DOI strings, quotations, or URLs.

OUTPUT
Return JSON only:
{
  "candidates": [{
    "visual_element": "...",
    "annotation": "one-sentence summary of the documented change",
    "periods": [{
      "period": "...", "year_start": 1840, "year_end": 1899,
      "context": "...", "meaning": "...",
      "search_terms": ["...", "..."]
    }],
    "topic_search_terms": ["...", "..."]
  }]
}

RULES
- No period may begin before 1840.
- Prefer visible practices, objects, clothing, symbols, spaces, and
  activities that can be rendered and inspected in a still photograph.
- If a proposed change cannot be stated without speculation, omit it.
- Search terms must target evidence, not merely repeat the topic title.
```

Listing 2: Librarian evidence-assessment prompt.

```
SYSTEM ROLE
You are the Librarian agent for the VDBench visual diachrony dataset. You receive
one candidate topic and documents retrieved for its claims. Your task is
to attach evidence, not to make the candidate sound convincing.

INPUT
```

```
- candidate topic JSON: {{candidate}}
- retrieved documents with title, author, date, URL, and text: {{documents}}

TASK
1. Match documents to the period or transition they actually support.
2. Retain only identifiable works with a locatable DOI or direct URL.
3. Reject generic search-result URLs, mismatched works, and sources that
   mention the element but do not support its historical change.
4. Mark every unsupported period or transition explicitly.
5. Never manufacture bibliographic metadata or repair a claim by guessing.

OUTPUT
Return JSON only:
{
  "visual_element": "...",
  "periods": [{
    "period": "...", "supporting_source_ids": ["src:0"],
    "evidence_note": "what the retained source supports",
    "supported": true
  }],
  "sources": [{
    "source_id": "src:0", "citation": "...", "doi": null,
    "url": "https://direct-work-or-record", "verified": false
  }],
  "unsupported_claims": [],
  "recommendation": "retain | revise | reject"
}

RULES
- A document is supporting evidence only when its accessible content or
  authoritative metadata can be inspected.
- Keep uncertainty visible. Unsupported is preferable to fabricated.
- Final verified status is assigned later by the source-audit pass.
```

Listing 3: Per-image specification prompt. Placeholders are filled with the reviewed period and the drawn generation parameters.

```
You are a visual-culture historian writing one grounded text-to-image specification for ONE
     photograph in ONE period. Your job is to describe visible material evidence, not to
     invent camera lore, named archives, photographer biographies, or unsupported technical
     backstory.

## TASK

Produce ONE specification for ONE photograph documenting `{{visual_element}}` at the period
     block below. The photograph must read as a real archival document - taken by a real
     person who was actually in the scene - and it must visibly carry the period's cultural
     meaning of the element through the world inside the frame.

You will see ONLY this one period. Reason from its `meaning` and `context` fields directly.
     The signifiers you choose are your interpretive work; that interpretation is what
     downstream analysis studies.

## OUTPUT - READ THIS FIRST

Your entire response is a single JSON object.
- First character: `{`
- Last character: `}`
- No markdown code fences. No `\`\`\`json`. No prose before the object. No prose after.
- Use double-quoted strings. Escape inner double quotes as `\"`.
- Every field listed in OUTPUT CONTRACT must be present.

## INPUT

### Entry (constant across the series)
- visual_element: {{visual_element}}
- shift_type: {{shift_type}}
- annotation: {{annotation}}

### Period - {{period_label}} ({{ys}}-{{ye}})
- year_start: {{ys}}
- year_end: {{ye}}
- context: {{context}}
- meaning: {{meaning}}
- aspect_ratio: {{aspect_ratio}}
- scene: {{scene}}
- activity: {{activity}}
- trend: {{trend}}
- style_anchors: {{style_anchors}}
```

```
- required_techniques: {{required_techniques}}

## GUIDANCE

The blocks below are reference material. Use them to shape the specification. Do NOT copy them verbatim into the output.

### Prose style
Write in documentary plainness: concrete physical facts, specific materials, light direction, period-relevant visible artifacts, and what hands are doing.

Match the prose to `style_anchors` when those are provided. When anchors are absent, do not infer capture medium, camera technology, archive, agency, photographer identity, or image-processing style from the date alone. Describe only visible scene content grounded in the supplied period, context, meaning, scene, activity, trend, and required techniques.

Real archival prose favours specific physical facts ("a cracked Bakelite handle", "off-white starched collar tinged with sweat at the neck", "fluorescent tube humming overhead", "rusted hinge with a pale paint smear") over abstract aesthetic adjectives. Concreteness beats decoration; a single specific noun is worth ten ambient mood words.

### Embodiment of meaning
The cultural meaning of `{{visual_element}}` in `{{period_label}}` must be CARRIED by the world inside the frame. Carry it through:

- garment fit, fabric, condition, and how it sits on the body
- what hands are doing - gripping, fixing, demonstrating, tucking, pointing, hiding, idle
- the orientation of bodies relative to each other and to the visual element
- the wear pattern on tools, surfaces, and the element itself
- the contents of the room - what is on the table, the walls, the shelf, the floor
- who is present and who is absent - class register, gender register, age register
- the architecture and decade-correct interior or exterior
- light quality and time of day appropriate to the scene

The cultural concept must be inferable from the arrangement of bodies, objects, surfaces, clothing, tools, and architecture. The strongest photographs let the viewer infer the meaning from concrete visible evidence.

When `trend` is non-empty, translate it into its photographed physical mechanism. A dance trend appears through posture, foot placement, rhythm, and onlookers. A toy or game trend appears through the object being handled and the social arrangement around it. A food or consumer fad appears through packaging, crumbs, containers, queues, counters, hands, and ordinary mess. A consumer-device trend appears through objects being handled, cables, furniture, and body posture. The trend is part of the scene's behavior, not a separate explanatory layer.

### Realism cues
Apply every cue listed in `required_techniques` to long_description, phrased positively. Describe what is present in the frame as concrete visual evidence.

Reference catalogue (for understanding the cues - do not pick from it yourself; the picks are pre-sampled in `required_techniques`):

- Framing: subject in the right or left third with empty space on the opposite side, horizon tilted a few degrees, a limb cropped at the frame edge, a foreground obstruction (a shoulder, a banister, a branch), a frame tilted slightly downward.
- Focus: focal plane on the background instead of the subject, motion blur on a hand or face, shallow depth of field accidentally resting on a secondary element.
- Light: harsh midday sun across half the face with skin detail intact, single-source flash with specular highlights and preserved fabric texture, window backlight silhouetting the subject, lens flare from a bare bulb.
- Surface and image artifacts: small occlusions, glare, reflection, weather marks, dust on nearby surfaces, worn edges, compression-like softness, or other visible imperfections only when they are requested in `required_techniques` or supported by `style_anchors`.
- Environment: bystanders entering the frame from the side, an era-appropriate vehicle in the background, weather (drizzle, wind-blown hair), clutter on a nearby surface, a cable or wire crossing the shot.

### Evidence boundary
Do not add capture-media claims or publication-history claims unless they are present in the input. This includes camera bodies, lenses, film stocks, sensors, named archives, photo agencies, photographer biographies, and assertions that an image is monochrome, color, digital, scanned, printed, or processed in a specific way.

If `style_anchors` provides concrete visual evidence, use that evidence as the style prior. If `style_anchors` is empty, leave capture medium and rendering technology unspecified and spend the description on visible people, objects, materials, spatial relations, activity, trend traces, and period-plausible scene content.

### Style anchors
```

```
When `style_anchors` is non-empty, treat the listed descriptions as the style prior for this period: visible palette, framing, light quality, surface texture, and level of formality. The long_description should sit comfortably inside the distribution implied by those anchors. They are evidence for what real documentation of this subject in this period looks like; let them override any generic decade assumption.

When `style_anchors` is empty, do not infer unsupported capture technology, publication channel, archive source, or image-processing style.

## INTERNAL PROCESS

Reason through the following before writing long_description. Use your reasoning channel; do NOT emit any of it into the JSON.

1. Read `meaning` and `context`. Hold the meaning as the interpretive target - never restate it. The image must SHOW the meaning, not narrate it.
2. Read `style_anchors`. What is the visual register of real documentation for this subject in this period?
3. Name the single strongest visual signifier - an object, posture, garment, spatial arrangement, gesture - that will anchor this photograph and visibly carry the meaning.
4. Pick `point_year` inside [{{ys}}, {{ye}}]. The opening sentence of long_description names this exact year.
5. Pick `composition_strategy` and `lighting` as short visible descriptions.
6. Do not invent a photographer persona or camera/lens backstory. If a human point of view matters, express it only through framing, distance, and what is visible in the frame.
7. Draft long_description in documentary plainness. Verify every cue in `required_techniques` is present, phrased positively.
8. Confirm the exact `composition_strategy` string appears verbatim in long_description.
9. Count the words. If under {{long_desc_word_min}}, add concrete visual facts - additional objects, surface textures, specific colours and materials, the quality of light on specific surfaces - until the count is inside [{{long_desc_word_min}}, {{long_desc_word_max}}]. Do not pad with abstract adjectives, gear trivia, or institutional backstory; pad with physical facts.

## HARD CONSTRAINTS

- Historical precision is paramount. Every garment, vehicle, storefront edge, hairstyle, and device must be plausible for {{period_label}}. If unsure, omit rather than invent.
- The visual element ({{visual_element}}) and its meaning are the primary subject. Scene, activity, and trend are scaffolding, never the subject.
- Cultural concepts (the meaning shift, the trend, the activity, the scene) are EMBODIED in the world (per Embodiment guidance) - through clothing, posture, gesture, objects, wear pattern, and architecture.
- If `trend` is non-empty, long_description must include concrete photographed traces of that trend's physical mechanism: bodies moving, objects handled, food or packaging used, devices operated, crowds arranged, or era-specific consumer artifacts in use.
- Scene + activity coherence: scene_selection and activity_selection are structural inputs you must echo verbatim in those JSON fields. INSIDE long_description, you have permission to interpret them naturally. If they feel non-overlapping (e.g. scene='desk' + activity='cycling'), do NOT depict the activity literally inside the scene noun. Instead, photograph the natural intersection - the subject pausing on the way, leaning a bicycle near the desk, mid-transition between locations. The scene is WHERE the photograph is taken near; the activity is WHAT is happening in that moment. A real archival photograph captures a believable moment, never a staged literal collision of nouns.
- `scene_selection` MUST be the exact string `{{scene}}`.
- `activity_selection` MUST be the exact string `{{activity}}`.
- `trend_selection` MUST be the exact string `{{trend}}` if a non-empty trend was provided; otherwise `trend_selection` MUST be `null`.
- `point_year` MUST be an integer inside [{{ys}}, {{ye}}] inclusive. The opening sentence of long_description must name this exact year.
- `aspect_ratio` MUST equal "{{aspect_ratio}}".
- `period` MUST equal "{{period_label}}".
- `year_start` MUST equal {{ys}} and `year_end` MUST equal {{ye}}.
- long_description MUST be {{long_desc_word_min}}-{{long_desc_word_max}} words. This is a HARD minimum - descriptions under {{long_desc_word_min}} words will be rejected. Use flowing present-tense third-person prose describing only what is visible inside the frame. No meta-commentary ("this image shows", "this photograph captures", "the era of"). Every sentence adds a new visual fact the image model can render.
- long_description must START with a sentence that names the subject, the action, and the specific point_year.
- long_description must contain the full `composition_strategy` string verbatim.
- long_description must contain every cue listed in `required_techniques`, phrased positively.
- People in the frame retain visible skin and fabric texture; bright surfaces stay off-white with detail rather than chalky textureless white.
- `description` is a one-to-two-sentence general caption of what is visible in the frame. It serves as the image caption.

## OUTPUT CONTRACT
```

```
Return a single JSON object with these fields. Unknown or inapplicable values are `null` (not omitted, not empty string).

{
  "period": "{{period_label}}",
  "year_start": {{ys}},
  "year_end": {{ye}},
  "point_year": <int inside [{{ys}}, {{ye}}] - the exact year this photograph is taken; named in the opening sentence of long_description>,
  "aspect_ratio": "{{aspect_ratio}}",
  "scene_selection": "{{scene}}",
  "activity_selection": "{{activity}}",
  "trend_selection": <"{{trend}}" if a non-empty trend was provided, otherwise null>,
  "composition_strategy": "<short phrase, e.g. 'wide-establishing low-angle' or 'tight over-shoulder crop'. Appears verbatim in long_description.>",
  "lighting": "<type + 1 sentence: direction, quality, color temperature>",
  "description": "<concise general caption - one or two sentences - of what is visible in the frame>",
  "long_description": "<{{long_desc_word_min}}-{{long_desc_word_max}} word present-tense third-person prose describing only what is inside the frame. Opens with point_year. Contains: (a) every cue from required_techniques, phrased positively; (b) the composition_strategy string verbatim>"
}

Notes on fields the downstream pipeline assembles itself (you do NOT produce these):
- where.places365_category - looked up from sources/places365_lookup.json keyed on scene_selection.
- trend.description - looked up from sources/filtered_trends.json keyed on trend_selection.
- The trajectory-level interpretive meaning is copied directly from the collection.

CRITICAL: long_description IS the text-to-image prompt. It is sent DIRECTLY to the image model with NO truncation or summarization. Every word matters. Concrete physical detail beats decorative adjectives. Hitting the {{long_desc_word_min}}-{{long_desc_word_max}} word range is NOT optional. Vague generic descriptions waste the word budget - every sentence should add a new visual fact the image model can render.

## FINAL REMINDER

Start your response with `{`. End it with `}`. One object. Every field present. No prose around it. No markdown fences.
```

Listing 4: The same template after substitution: the complete prompt supplied to Nano Banana Fast for *Radio*, period 1940s–1950s, with the model-selected year 1955.

```
Render as an authentic 1940s-1950s photograph: PREDOMINANTLY BLACK-AND-WHITE press/documentary tonality with coarse film grain and flat mid-tones; only occasionally muted early Kodachrome/Agfacolor with desaturated, slightly faded, yellow-green-shifted hues (NEVER bright or saturated). Low dynamic range, plain lighting, and soft detail. This is an ORDINARY period photograph with the technical limitations of its era: NOT polished, NOT cinematic, NOT studio-lit, NOT high-dynamic-range, NOT a modern digital render. Flatter contrast, imperfect exposure, plain available light, mild softness and blur, and visible film grain are expected. Candid documentary realism: the subject is mid-action, attention on the task, another person, or an off-frame event; composition is informal and slightly off-balance, as if grabbed in passing rather than arranged. Mild archival handling wear: small dust marks, softened fine detail, slight edge wear, and subtle uneven tone while preserving faces, clothing, and the primary subject. Core visual brief: diachronic subject = Radio; period = 1940s-1950s. Carry the brief through photographed people, objects, materials, behavior, and setting. In 1955, a young copy runner holds a pocket transistor radio toward three newsroom workers as they stop marking wire copy and listen. The close over-desk diagonal places the small cream-and-red receiver near the center, its perforated grille facing the group and its thin tuning dial pinched between the runner's wet thumb and forefinger. A narrow leather carrying strap hangs below his hand, darkened where it has rubbed against his palm. His short wool jacket carries beads of drizzle across the shoulders, and wind-blown hair or drizzle visible on clothing and nearby surfaces gives the interruption the disorder of someone who has just hurried in from the street.

The frame tilted slightly downward takes in the scarred oak desk, overlapping news sheets, blue editing pencil, steel scissors, ashtray, coffee rings, and a black desk telephone with its coiled cord crossing a typed page. One middle-aged editor in shirtsleeves leans away from his typewriter, pencil suspended above a paragraph. His loosened dark tie rests crookedly against an off-white cotton shirt whose creases, stitching, and faint dampness at the collar remain visible. A woman at the adjoining desk turns sideways in a fitted cardigan and plain calf-length skirt, one hand holding a fresh sheet from a wire-service printer while the other presses two fingers against the paper to keep it from curling. Behind her, a younger reporter in a narrow-lapelled jacket bends forward with both palms on the desk, his face oriented toward the portable set rather than the copy before him.
```

```
5 A larger wood-cased radio sits beneath a wall clock on a metal filing cabinet, its cloth
      speaker grille worn pale at the corners and its power cord descending toward a crowded
      socket. Pinned maps, pencilled telephone numbers, assignment slips, and rows of clipped
       bulletins cover the wall. A strip of paper trails from the wire machine into a shallow
       basket, with several lines circled heavily in pencil. Through the rain-streaked window
      , a rounded postwar sedan stands at the curb; its dashboard radio and horizontal dial
      are visible through the open driver's door, where a hat rests on the front seat.
6
7 Cool window light falls from the left across damp wool and the portable radio's ridged
      plastic case. Green-shaded desk lamps make warmer islands over typing paper and
      preserve skin pores, knuckle creases, and fabric weave. Bright collars and paper remain
       off-white with folds, smudges, and typed detail. The shallow depth of field
      accidentally resting on a secondary element makes the nearest coffee cup and its
      chipped rim unusually crisp while the runner's sleeve and distant filing cabinets
      soften slightly. Mild archival handling wear with small dust marks and slight edge wear
       appears around the frame without concealing faces, hands, radio controls, or newsroom
      clutter. Avoid large areas of pure blown-out white.
```

D Dataset Validation Details

D.1 Review Protocol and Aggregation Rule

Validation runs in two stages with the same pool of reviewers. Stage 1, before any image exists, assesses whether a topic is a defensible diachronic topic and whether its periods are correctly drawn. Stage 2, after generation, assesses whether the rendered images are plausible for the topic and period they claim to depict. Reviewers are domain experts in the humanities and social sciences; each chose topics from the full audited list according to their own expertise rather than being assigned topics, so coverage is uneven by design.

Topic decisions use a strict majority over the reviews actually completed for that topic: approved when more than half are positive on both topic-level criteria, rejected when more than half are negative, and recorded as a tie when the counts are equal. Ties and rejections are both excluded from image generation. Because 37 of the 46 reviewed topics received exactly one review, the rule reduces to a single reviewer's decision for most topics; only nine topics were reviewed more than once. This is the concrete form of the review-depth limitation in Sec. 7.

D.2 Stage 1: Topic and Trajectory Review

Procedure. For each selected topic the interface showed its title, summary, complete period sequence, and verified sources. The expert first answered whether the topic exhibits genuine visible change over time and whether it represents a culturally important phenomenon; a review counted as positive only if both answers were positive. These judgements concern the topic as a documented phenomenon, not the reviewer's agreement with its social or political meanings. The expert then worked through every period description, rated agreement on a five-point scale, could edit start and end years, and could leave comments (Fig. 4).

Outcomes. Six experts completed 57 reviews of 46 distinct topics, 51 of which were positive on both criteria. Strict majority approved 41 topics, rejected three, and left two tied; only the 41 approved topics proceeded to generation. At period level the experts submitted 270 ratings for 221 distinct descriptions, with mean 4.02 and median 4.00 on the five-point scale. Of the nine topics reviewed more than once, six drew the same decision from every reviewer. *Camera* was approved on two positive reviews against one negative. *Telephone* and *Headphones* each received one positive and one negative review and were excluded as ties. *Television set*, *DNA double helix image*, and *Face filter / AR face effect* were rejected.

Expert feedback. Comments did more than decide inclusion; they changed period boundaries, narrowed geographic claims, and corrected accounts of what is visible in a given period (Tab. 10). Two criticisms recur and bear directly on Sec. 7. Several reviewers found trajectories framed from a Western, often specifically North American, perspective and asked for European and post-Soviet contexts. Others separated the depicted object from the photographic style of its era, warning that a topic can appear to change simply because photography changed.

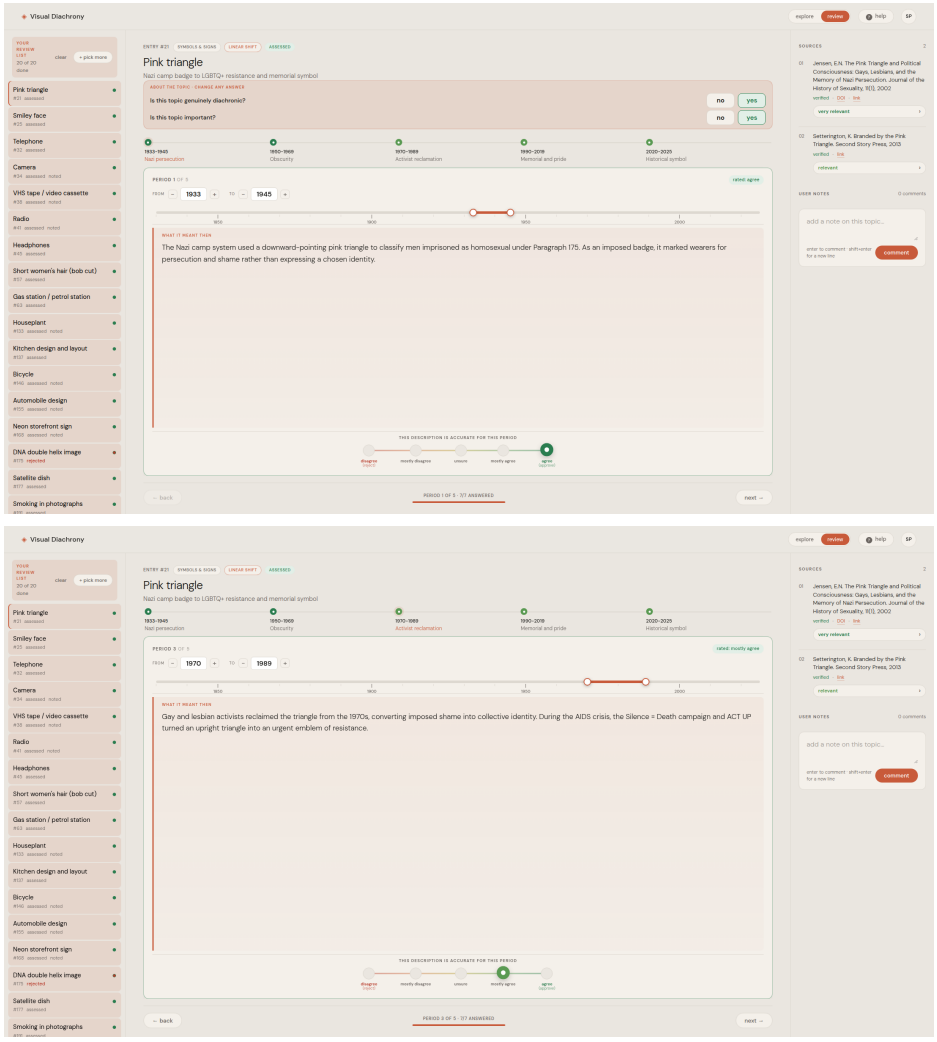

Figure 4: Stage 1 review interface. Top: the two topic-level questions, the trajectory, a period assessment, the verified sources, and the comment panel. Bottom: a later period with editable dates, its description, and the five-point agreement scale.

Table 10: Stage 1 expertise, review load, and principal revisions. The final column paraphrases the submitted free-text comments.

ID	Expertise	Topic reviews	Principal revisions
A	Gender studies and sociology	10	Strengthened the treatment of women's agency and of the political uses of debates about the hijab. Added racialised differences and differences between cisgender and transgender experiences to body-hair norms, qualified Western and US framing, and identified omitted transgender and intersex debates in women's sport.
B	Sociology and social anthropology	3	Distinguished devotional, material, ritual, and oppositional meanings of cross jewellery, and noted that these meanings continue to coexist rather than succeeding one another.
C	Media and communication studies	12	Added missing contexts and earlier dates for trousers and sneakers. Distinguished the athletic, youth, and anti-corporate meanings attached to hoodies, qualified interpretations of tattoos, and revised the account of brutalist architecture.
D	Media studies, literary studies, and digital humanities	6	Identified North American bias in the mall and advertising trajectories and added post-Soviet and European contexts. Requested amateur and pirate radio, and questioned the 1910 selfie boundary and the earliest podcast period.
E	Film archiving, media studies, and film theory	20	Separated the depicted object from the photographic style of its era for *Camera* and *VHS*, revised starting dates, and flagged geographic limits. Rejected the DNA helix as an insufficient visual marker, and questioned later neon-sign evidence and the visual distinctiveness of podcast-era radio.

Table 10: Stage 1 expertise, review load, and principal revisions (continued).

ID	Expertise	Topic reviews	Principal revisions
F	Migration, mobility, diaspora, borders, and identity	6	Qualified claims about corsets and physical freedom, differentiated forms and settings of head covering, and corrected *Camera* and *Radio* dates. Questioned broad or normative periods for bob cuts and face filters.

D.3 Stage 2: Image Plausibility Review

Procedure. Experts A, B, E, and F returned after generation. For each period they saw the same topic title, complete trajectory, and period descriptions as in Stage 1, together with a random sample of three of the 100 generated images. They were told explicitly that the images were AI-generated, and were asked to judge plausibility and correspondence to the stated topic and period on first inspection rather than to search for generation artifacts. Image labels, prompts, reference images, model settings, and the automated verification outputs stayed hidden (Fig. 5).

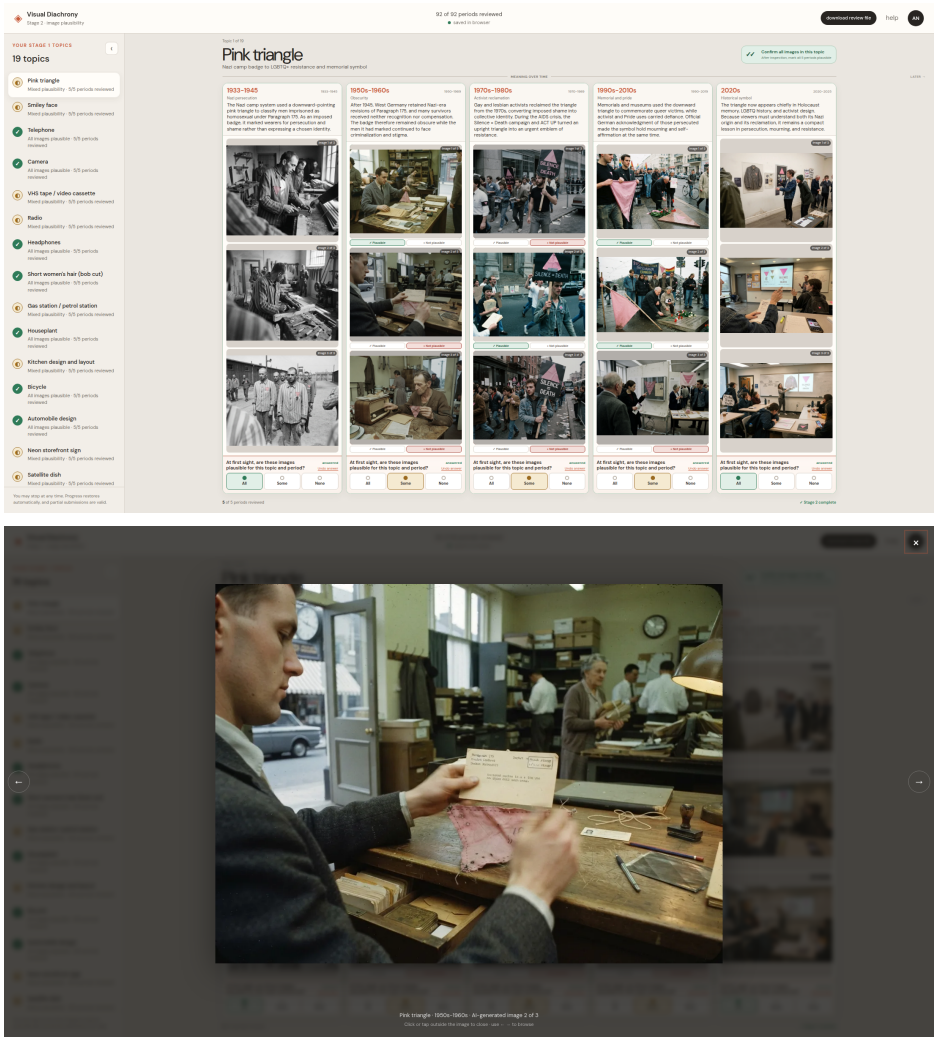

Figure 5: Stage 2 review interface. Top: the complete trajectory with three sampled images per period and the image-level plausibility controls. Bottom: the enlarged single-image view in which the judgement is recorded.

Outcomes. Of the 501 sampled images, 422 (84.2%) were judged plausible (Tab. 11), with per-expert rates between 71.7% and 93.8%. Five topics were reviewed by two experts, whose independently drawn samples contained 53 images in common; the two judgements matched for 38 of them

Table 11: Completed Stage 2 reviews by expert, covering 29 distinct topics. Per-expert rates are descriptive only: reviewers chose different topics and saw independently drawn samples, so the rows are not directly comparable.

	Topic reviews	Period reviews	Images	Plausible
Expert A	10	49	147	116 (78.9%)
Expert B	3	16	48	45 (93.8%)
Expert E	17	82	246	218 (88.6%)
Expert F	4	20	60	43 (71.7%)
Total	**34**	**167**	**501**	**422 (84.2%)**

(71.7%). We report this raw agreement but do not estimate a dataset-wide inter-rater coefficient from an overlap this small.

D.4 Scope of the Validation

The two stages certify different things, and neither certifies the historical accuracy of an individual image. Stage 1 establishes that every trajectory used for generation was judged defensible by at least one domain expert and that its boundaries and descriptions were reviewed and, where necessary, corrected; it covers all 41 generated topics. Stage 2 establishes that images drawn at random are usually plausible for the topic and period they are labelled with, but it covers 501 of 20,100 images, or 2.5% of the collection, so it bounds the overall error rate only loosely and says nothing about any specific image. The automated verification of App. B.3 does cover every image, but tests only whether the requested environment, activity, and optional trend are visible, not whether the scene is free of anachronism.

E Evaluation Details

Task 2 compares two groupings of the same images: the ground-truth periods $\mathcal{P} = (P_1, \ldots, P_J)$ and the predicted periods $\hat{\mathcal{P}} = (\hat{P}_1, \ldots, \hat{P}_{\hat{J}})$. Some metrics compare the groupings as wholes and need no correspondence between individual periods; others compare a predicted period with a specific ground-truth period and therefore require one. Table 12 states which metrics depend on that correspondence and how each treats a period with no counterpart.

Table 12: Dependence of each Task 2 metric on period alignment. "Unmatched periods" means ground-truth periods with no matched prediction, or predictions with no matched ground-truth period.

Output	Metric	Needs alignment	Treatment of unmatched periods
Image grouping	ARI, AMI	No	Accounted for implicitly: every image is in exactly one predicted group
Period count	$\lvert\hat{J} - J\rvert$	No	Counted directly
Period range	Temporal IoU	Yes	Excluded from the average; App. H.2 reports the zero-filled variant
Period title, description	BERTScore, MoverScore	Yes	Excluded from the average
Matched order	Kendall's τ	Yes	Excluded; needs at least two matched pairs and is reported over eligible topics only
Full sequence	Normalised Levenshtein	Yes, for encoding	Encoded as distinct placeholders, so they cost insertions or substitutions
Topic title, description	BERTScore, MoverScore	No	Not applicable: one prediction against one reference

Period-level quality metrics ask how good a period is *given* that it corresponds to a ground-truth period, so scoring an unmatched period would confound range accuracy with segmentation failure.

Missing and additional periods are instead the concern of period-count error, ARI, AMI, and the sequence distance. The *trajectory structure* rows of Tab. 4 therefore have to be read together; no single one summarises the others.

E.1 Period Alignment

Let $\mathcal{I}_j$ be the images in ground-truth period j and $\hat{\mathcal{I}}_k$ those assigned to predicted period k. Their agreement is the intersection over union of the two image sets,

$$A_{jk} = \frac{|\mathcal{I}_j \cap \hat{\mathcal{I}}_k|}{|\mathcal{I}_j \cup \hat{\mathcal{I}}_k|} \in [0, 1], \tag{1}$$

which is 1 when the two periods contain the same images and 0 when they share none. We then choose the set M of one-to-one pairs (j, k) maximising $\sum_{(j,k)\in M} A_{jk}$, using Hungarian matching (Kuhn, 1955). At most $\min(J, \hat{J})$ pairs can be formed, and pairs with $A_{jk} = 0$ are discarded, so M may be smaller still.

Matching jointly rather than greedily matters: a greedy rule that gives each prediction its own best ground-truth period can leave a strong pair unmatched because a weaker prediction claimed that period first, as App. E.4 shows. One alignment is computed per topic and reused for every period-level metric, so temporal IoU, the period texts, and Kendall's τ all refer to the same pairs. Alignment is a preprocessing step and not itself a metric, which is why it does not appear in Tab. 4.

For the sensitivity analysis in App. H.2 only, we also use a penalised aggregation that charges for unmatched periods explicitly,

$$Q_{\text{pen}} = \frac{\sum_{(j,k)\in M} q_{jk} + \big(\max(J, \hat{J}) - |M|\big)\, q_{\min}}{\max(J, \hat{J})}, \tag{2}$$

where q_{jk} is the value of the period-level metric on a matched pair and $q_{\min}$ is the lowest value that metric can take. Every period without a counterpart, on either side, is assigned $q_{\min}$. Table 4 uses the matched-pair average instead.

E.2 Temporal IoU

Period ranges are inclusive, so $R = [s, e]$ denotes the calendar years $Y(R) = \{s, s+1, \ldots, e\}$. For a matched pair,

$$\text{IoU}_{\text{year}}(R_j, \hat{R}_k) = \frac{|Y(R_j) \cap Y(\hat{R}_k)|}{|Y(R_j) \cup Y(\hat{R}_k)|}. \tag{3}$$

Identical ranges score 1 and disjoint ranges 0. For example, $[1900, 1919]$ and $[1910, 1929]$ each span 20 years; their 10-year intersection and 30-year union give $10/30 = 1/3$. Note that $\hat{R}_k$ is not predicted directly but derived from the years of the images assigned to period k (Sec. 5.1), so temporal IoU responds both to the image assignment and, in end-to-end evaluation, to Task 1 year error.

E.3 Sequence Distance

The sequence distance asks whether the predicted trajectory tells the same story in the same order, including periods that are missing or invented. Both sequences are taken in chronological order, which for the prediction is the order the task definition requires (Sec. 5.1). Each predicted period is encoded by the index of the ground-truth period it was matched to, and every unmatched prediction receives its own placeholder that matches no ground-truth index. This gives $\hat{\mathbf{o}}$ of length $\hat{J}$, compared against $\mathbf{o} = (1, 2, \ldots, J)$ through

$$d_{\text{Lev}} = \frac{\text{Lev}(\mathbf{o}, \hat{\mathbf{o}})}{\max(J, \hat{J})}, \tag{4}$$

where Lev is the minimum number of insertions, deletions, and substitutions turning one sequence into the other. A missing period costs an insertion, an invented period a deletion, and a misordered period a substitution, so the three failure modes are charged separately. For ground-truth

order $(1, 2, 3, 4)$, the incomplete prediction $(1, 2, 4)$ needs one insertion and scores $1/4$, while the misordered prediction $(2, 1, 3, 4)$ needs two substitutions and scores $2/4$.

E.4 A Worked Example

Take a topic with 100 images and $J = 2$ ground-truth periods, $\mathcal{I}_1 = \{1, \ldots, 50\}$ and $\mathcal{I}_2 = \{51, \ldots, 100\}$, images numbered chronologically. A method predicts $\hat{J} = 3$ periods, $\hat{\mathcal{I}}_1 = \{1, \ldots, 30\}$, $\hat{\mathcal{I}}_2 = \{31, \ldots, 60\}$, and $\hat{\mathcal{I}}_3 = \{61, \ldots, 100\}$: it splits the first period and shifts the boundary. Equation (1) gives the agreement matrix in Tab. 13.

Table 13: Agreement matrix A_{jk} for the worked example. Bold entries are the pairs selected by Hungarian matching.

	$\hat{P}_1$ (1–30)	$\hat{P}_2$ (31–60)	$\hat{P}_3$ (61–100)
P_1 (1–50)	**0.60**	0.33	0.00
P_2 (51–100)	0.00	0.14	**0.80**

A greedy rule walking through the predictions in order gives $\hat{P}_1$ its best period P_1 (0.60), then $\hat{P}_2$ the only remaining period P_2 (0.14), and leaves $\hat{P}_3$ unmatched despite it being the strongest pair in the matrix: total 0.74. Hungarian matching returns $M = \{(1, 1), (2, 3)\}$ with total 1.40, leaving $\hat{P}_2$ unmatched. Temporal IoU and the period texts are then averaged over the two matched pairs, and $\hat{P}_2$ contributes nothing to them. Period-count error is $|3 - 2| = 1$. Kendall's τ uses the two matched pairs, whose orders agree, giving $\tau = 1$. The sequence distance encodes the prediction as $\hat{\mathbf{o}} = (1, \times, 2)$ against $\mathbf{o} = (1, 2)$, where $\times$ is the placeholder for $\hat{P}_2$, so one deletion suffices and $d_{\text{Lev}} = 1/\max(2, 3) = 0.33$. ARI and AMI act on the image partitions directly and need no alignment. The example is also why the structural metrics must be read together: a method can score $\tau = 1$ on order while inventing a period, and only the count error and the sequence distance reveal it.

F Baseline Implementation Details

F.1 Common Protocol

Both baselines use frozen weights, receive only the inputs the task definition allows, and use no VD-Bench data for training, fine-tuning, or prompt selection. File names and record identifiers reveal no period information; Task 2 annotations and construction metadata are visible only to the evaluation code; and the true topic name, period count, and period ranges are withheld in Task 2. Ground-truth Task 1 fields reach a baseline only in the GT setting, where the task definition provides them. In the Pred. setting a baseline consumes its own Task 1 output and nothing else.

Table 14 maps every benchmark output to the component that produces it. Open-MLLM routes all eight outputs through one model, while TrajEx assigns each output to a specialised component and derives the period structure algorithmically rather than by generation.

F.2 Open-MLLM

Qwen3-VL-32B-Instruct (Qwen Team, 2025) predicts the year, environment, activity, and caption in four separate calls per image, so no Task 1 output conditions another. For the two closed vocabularies the model returns a label together with its index; a call whose label and index do not correspond is repeated, which removes ambiguous parses without correcting the model's choice.

Task 2 is text-only. The model receives one record per image containing the year, environment, activity, and caption, and never sees pixels again. It proposes the hidden topic with a title and description, proposes a sequence of periods with their own titles and descriptions, and assigns every image record to exactly one period. The evaluated ranges $\hat{R}_k$ are not generated but derived from the Task 1 years supplied in that setting (Sec. 5.1), so the model controls the range only through its image assignment.

Table 14: Which component produces which benchmark output.

Output	OPEN-MLLM	TRAJEX
Task 1, per image		
Year $\hat{t}_i$	Qwen3-VL-32B-Instruct	Temporal projection of Tekaya et al. (2025) on a frozen EVA02-CLIP-L-14-336 encoder (Sun et al., 2023)
Environment $\hat{e}_i$	Qwen3-VL-32B-Instruct	EVA-CLIP zero-shot similarity
Activity $\hat{a}_i$	Qwen3-VL-32B-Instruct	EVA-CLIP zero-shot similarity
Caption $\hat{c}_i$	Qwen3-VL-32B-Instruct	Qwen3-VL-8B-Instruct
Task 2, per topic		
Period count $\hat{J}$	Proposed by the model	Change-point penalty chosen for stability across subsamples
Image assignment $\hat{\mathcal{I}}_k$	Proposed by the model	Contiguous segments of the year-ordered images, then merging of undersized segments using DINOv3 features (Siméoni et al., 2025)
Period title, description	Qwen3-VL-32B-Instruct	Qwen3-VL-8B-Instruct, from five medoid images per period
Topic title, description	Qwen3-VL-32B-Instruct	Qwen3-VL-8B-Instruct

F.3 TRAJEX

TRAJEX (*Trajectory Explorer*) replaces generation with an explicit segmentation stage. For Task 2 it sorts a topic's images by their supplied Task 1 year and applies dynamic-programming change-point detection (Truong et al., 2020) to that one-dimensional sequence, so the initial periods are contiguous in time by construction. The penalty controlling the number of change points is selected without ground truth, by keeping the result that stays stable when the input is subsampled. DINOv3 visual features (Siméoni et al., 2025) and text embeddings of the supplied captions and labels then merge segments too small to describe and select five medoid images per surviving period. Qwen3-VL-8B-Instruct writes the period and topic texts from those medoids together with the captions and label counts.

Two consequences of this design run through the results. Because segmentation uses the year field alone, TRAJEX inherits Task 1 year error directly (App. H.7) and collapses when the year signal is destroyed (App. H.9). Because its segments are contiguous in time, it can merge neighbouring periods but cannot interleave images from distant ones, which is the opposite failure profile to OPEN-MLLM. App. H.8 tests whether feature-space clustering is a viable substitute for the temporal segmentation.

G TOPIC-LEVEL RESULTS AND EXAMPLES

The aggregate Task 2 scores in Tab. 4 are macro-averages over 41 topics and hide substantial variation.

G.1 VARIATION ACROSS TOPICS

With GT input, TRAJEX reaches ARI of at least 0.9 on 22 topics and falls below 0.5 on seven, while OPEN-MLLM leads on nine topics (Fig. 6). With Pred. input the two macro-averages are close (0.41 and 0.46 in Tab. 4), yet OPEN-MLLM leads on 19 of the 41 topics. Nearly equal averages do not mean the methods behave alike.

Figure 7 separates the two ways a method can group images badly: predicting the wrong number of periods, and assigning images to the wrong ones. At zero count error the vertical position reflects assignment alone.

The two baselines fail differently. With GT input TRAJEX returns the correct count for 21 topics and averages ARI 0.99 on exactly those; on the seven topics where it returns only two periods, mean ARI falls to 0.37. With Pred. input it under-segments 33 topics and returns two periods for 20 of them. OPEN-MLLM returns the correct count for 21 topics with GT input and 19 with Pred. input, but its

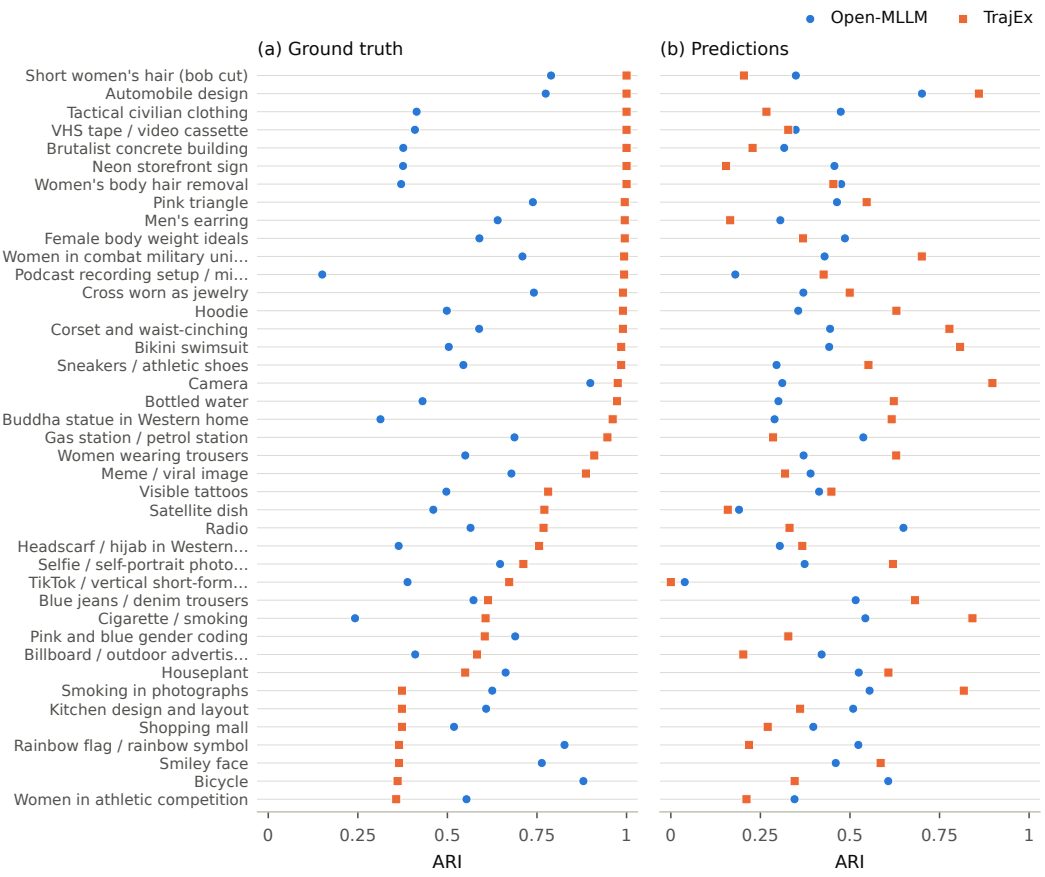

Figure 6: Task 2 image-grouping ARI per topic with GT Task 1 values (a) and with each baseline's own predictions (b). Both panels use the topic order defined by TRAJEX's score in (a).

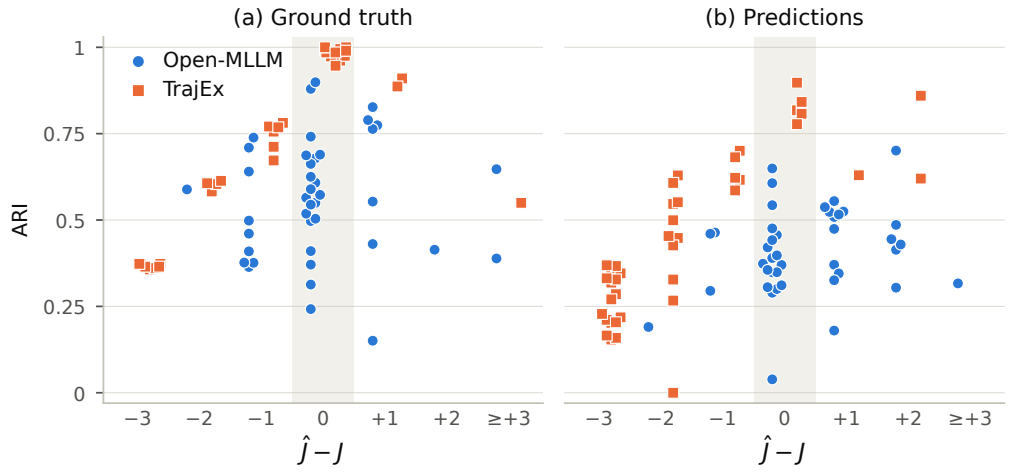

Figure 7: Task 2 ARI against the signed period-count difference $\hat{J} - J$, with GT Task 1 values (a) and predictions (b). Positive values indicate over-segmentation; differences of $+3$ or more share the rightmost bin.

mean ARI on those topics is only 0.58 and 0.39. So TRAJEX groups images almost perfectly once it has the right number of periods, making count selection its binding constraint, while OPEN-MLLM more often finds a plausible count but mixes images across the periods it proposes. App. H.11 confirms this by supplying the true count directly.

G.2 ASSIGNMENT PATTERNS

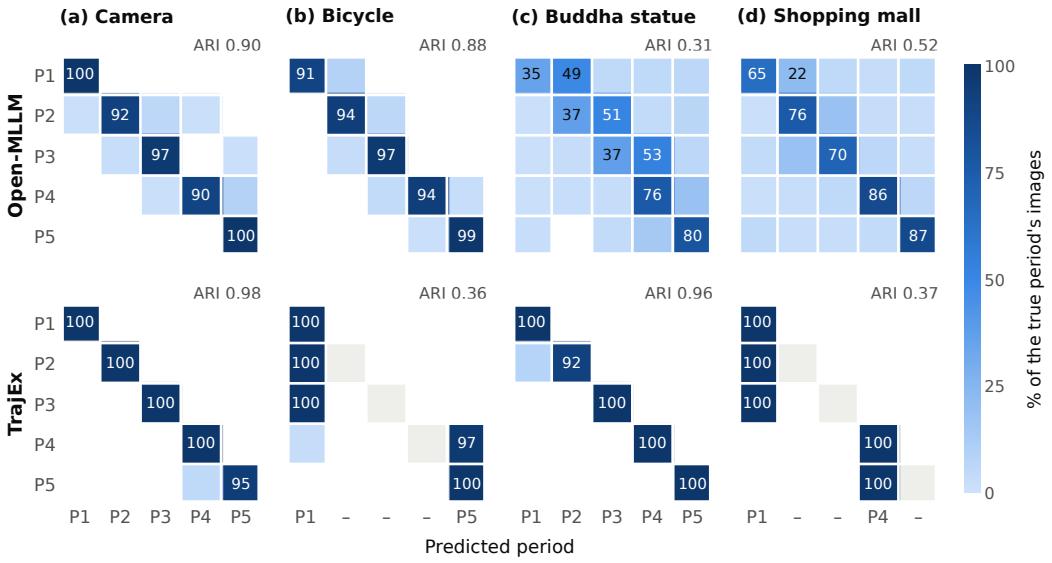

Figure 8: Image assignments with GT input for the four gallery topics. Rows are ground-truth periods, columns are the predicted periods matched to them, and each cell gives the percentage of a row's images assigned to that column. Values below 20% are unlabelled, and a minus sign marks a ground-truth period with no matched prediction. Correct assignments lie on the diagonal. The panels show both baselines succeeding (a), TRAJEX merging periods (b), OPEN-MLLM mixing periods (c), and both failing (d).

The four gallery topics in Fig. 8 were chosen by rules fixed before inspection. Among topics on which both methods do well, *Camera* maximises the lower of the two ARI scores. *Bicycle* has the largest OPEN-MLLM advantage among topics that TRAJEX under-segments. *Buddha statue in Western home* has the largest TRAJEX advantage among topics for which OPEN-MLLM predicts the correct count. Among the remaining topics, *Shopping mall* has the lowest score from the better baseline. The patterns are consistent across panels: TRAJEX errors are whole periods merged into a neighbour, so its off-diagonal mass sits in adjacent rows, whereas OPEN-MLLM spreads a single ground-truth period across several predicted ones.

G.3 MATCHED PERIOD EXAMPLES

Four topics were selected to cover four distinct outcomes rather than four assignment patterns, so the set differs from the gallery in App. G.2, in which *Camera* and *Bicycle* also appear. Table 15 gives all four; the two tables that follow print their predicted trajectories in full.

In Tabs. 16 and 17 a row is a ground-truth period, and a cell gives the title and derived range of the prediction matched to it under App. E.1; ranges span the earliest and latest supplied year among the assigned images. "No match" marks a ground-truth period that no prediction was matched to. Predicted periods matched to no ground-truth period have no row and are not shown, but they are still counted by the period-count error and the sequence distance (Tab. 12).

Camera. Both baselines match all five periods with GT input, at ARI 0.90 and 0.98 (Tab. 16). TRAJEX holds at 0.90 with Pred. input, whereas OPEN-MLLM falls to 0.31 while still returning five periods: its Task 1 errors damage the image assignment, not the count. The wide OPEN-MLLM ranges, such as 1963–2025 for the phone-camera period, follow from the same cause, since a period that collects images from the wrong decades inherits their years.

Table 15: The four example topics and why each was selected. Scores are ARI / temporal IoU; J is the number of ground-truth periods.

			OPEN-MLLM		TRAJEX	
Topic	J	**Selected as**	**GT**	**Pred.**	**GT**	**Pred.**
Camera	5	Both succeed	0.90 / 0.46	0.31 / 0.45	0.98 / 0.65	0.90 / 0.60
Bicycle	5	OPEN-MLLM leads	0.88 / 0.62	0.61 / 0.36	0.36 / 0.26	0.35 / 0.17
Women wearing trousers	6	Main-paper example	0.55 / 0.31	0.37 / 0.19	0.91 / 0.69	0.63 / 0.69
Podcast recording setup / microphone on desk	4	Largest TRAJEX GT lead	0.15 / 0.40	0.18 / 0.29	0.99 / 0.88	0.43 / 0.50

Table 16: Matched predicted periods for *Camera*.

Ground truth	**Open-MLLM GT**	**Open-MLLM Pred.**	**TrajEx GT**	**TrajEx Pred.**
Specialist tool 1840–1889	Wet-Plate Specialist Era 1848–1903	Portable Camera and Amateur Photography 1880–1905	Wet-Plate Photography 1848–1887	The Photographic Studio Era 1849–1897
Democratization 1888–1939	Box Camera Democratization 1896–1964	Rise of the Folding and Box Cameras 1890–1930	The Amateur Photographer 1896–1938	The Amateur Photographer 1899–1948
SLR era 1940–2009	35mm SLR Mainstreaming 1912–1989	Mid-Century Camera and Social Documentation 1920–2023	The Everyday Photographer 1959–1978	The Amateur Photographer 1950–1983
Digital revolution 1980–2009	Digital and Disposable Transition 1913–2008	*No match*	Disposable and Digital Cameras 1987–2014	Digital Camera Enthusiasts 1986–2014
Phone camera era 2010–2025	Smartphone Camera Ubiquity 1963–2025	Digital Camera and Personal Technology 1955–2023	Smartphone as Everyday Lens 2015–2025	Smartphone Documentation 2015–2021
ARI / temporal IoU	0.90 / 0.46	0.31 / 0.45	0.98 / 0.65	0.90 / 0.60

Table 17: Matched predicted periods for *Women wearing trousers*, the topic the main paper calls *women in trousers*.

Ground truth	**Open-MLLM GT**	**Open-MLLM Pred.**	**TrajEx GT**	**TrajEx Pred.**
Dress reform 1850–1899	Practical Transition 1871–2018	Practical Workwear 1870–1975	Women in Turkish Trousers 1851–1904	Working Women in Early Industrial Era 1850–1929
Suffrage era 1900–1919	Wartime Normalisation 1906–2018	Professional and Civic Wear 1900–2015	Women in Practical Activewear 1906–1919	*No match*
Interwar fashion 1920–1939	*No match*	Fashionable and Everyday Wear 1880–2010	Sailor Trousers in Everyday Life 1923–1936	*No match*
WWII practicality 1940–1959	Postwar Integration 1926–2007	Hidden Undergarment 1850–1975	Women in Utility Clothing 1937–1958	The Everyday Woman in Mid-20th Century Work and Service 1931–1955
Liberation era 1960–1979	Cultural Mainstream 1968–1996	Mainstream and Diverse Use 1890–2006	The Working Woman in the 1970s 1964–1979	Women in Transition 1962–1981
Full acceptance 1980–2025	Contemporary Ubiquity 1994–2025	Contemporary and Versatile Wear 1850–2023	Urban Domestic Worker 2015–2025	Domestic and Civic Labour 1985–2020
ARI / temporal IoU	0.55 / 0.31	0.37 / 0.19	0.91 / 0.69	0.63 / 0.69

***Women wearing trousers*.** Neither baseline succeeds here (Tab. 17). With GT input TRAJEX reaches 0.91 and OPEN-MLLM 0.55, and both fall with Pred. input. The topic shows why grouping and range overlap are reported separately: TRAJEX keeps temporal IoU at 0.69 with Pred. input even with two ground-truth periods unmatched, because the periods it does return are correctly placed in time, and the missing ones are charged by the count error and the sequence distance instead. The period titles also differ in kind. TRAJEX names visual configurations ("Sailor Trousers in Everyday Life"), OPEN-MLLM names historical processes ("Wartime Normalisation"), and the references are process names such as "Suffrage era", which is consistent with the period-text scores in Tab. 4 favouring OPEN-MLLM.

***Bicycle* and *Podcast*.** The other two selections show the same two failure modes in extreme form; their scores are in Tab. 15 and *Bicycle*'s assignment matrix is panel (b) of Fig. 8. On *Bicycle*, OPEN-MLLM recovers all five periods with GT input and keeps the count with Pred. input, while TRAJEX returns two periods in both settings, collapsing the first four ground-truth periods into one 1867–2001 group and leaving three without a counterpart; its matched pairs are reasonable, but most of the trajectory is absent, which is why period-count error and the sequence distance are reported beside ARI. *Podcast recording setup / microphone on desk* is the opposite case: its first period runs 1920–2004 while the remaining three fall between 2004 and 2025, and with GT years TRAJEX separates them almost exactly (ARI 0.99, temporal IoU 0.88) because contiguous segmentation of exact years resolves boundaries a few years apart. With predicted years it loses the two middle periods, whose nine- and six-year spans are comparable to a year MAE near five (Tab. 4), while OPEN-MLLM produces broad overlapping ranges in both settings and scores 0.15 and 0.18. Topics whose periods are no longer than the achievable dating error are the hardest case in VDBench.

H ABLATION AND DIAGNOSTIC STUDIES

H.1 WHAT THE STUDIES TEST AND HOW THEY ARE REPORTED

Table 18: The diagnostic studies and the question each one answers.

Study	Question
Do the metrics measure what they should?	
App. H.2	Does averaging temporal IoU over matched pairs flatter a method that omits periods?
App. H.3	How much can BERTScore and MoverScore actually resolve?
App. H.4	Which differences in Tab. 4 are consistent across topics?
What do the baselines rely on, and what can be solved without them?	
App. H.5	How well do non-visual rules do, and is equal period size a shortcut?
App. H.6	Is Task 1 dating tied to one image cue, one era, or model size?
App. H.7	Which Task 1 field does each baseline's Task 2 grouping depend on?
App. H.8	Is feature-space clustering a viable substitute for temporal segmentation?
App. H.9	What survives when the year signal or the image content is destroyed?
App. H.10	Is OPEN-MLLM reading caption content or just the dates inside captions?
App. H.11	How much of the remaining error is period-count selection?

Common protocol. Unless a control changes the collection, Task 2 diagnostics use all 41 topics, 201 periods, and 20,100 images, with scores computed per topic and then macro-averaged; temporal IoU and period-text scores use matched periods only. The Task 1 reference rules, the era analysis, and the uncertainty estimates also use all 20,100 images. The image perturbations and the model-size control use one fixed subset of 25 images per topic, 1,025 in total, so their absolute scores differ slightly from the full-collection values in Tab. 4 and should be compared only within that study.

Uncertainty. Brackets give a 95% confidence interval for the mean across topics, obtained by resampling the 41 topics 10,000 times with all images of a topic kept together. Paired comparisons use the same resampled topics on both sides. Following the main paper, we call a paired difference *consistent across topics* when $p < 0.05$; an interval for a difference that contains zero establishes no advantage, and equally does not establish equivalence. The resampling unit is the topic rather than the image, because images within a topic are not independent.

Multiple comparisons. These studies report 55 paired decisions at $p < 0.05$, and those decisions are not corrected for multiplicity: across a family that size the probability of at least one false positive under the global null is about 0.94 if the tests were independent, and they are not independent, since many reuse the same runs. We therefore distinguish two classes of result and rely only on the first. The conclusions we draw rest on effects whose intervals sit far from zero: the 0.445 and 0.127 ARI drops under content shuffling (App. H.9), the 0.337 [0.238, 0.431] and 0.115 [0.084, 0.148] ARI reductions from replacing a Task 1 field (App. H.7), the 0.158 ARI and 0.104 temporal-IoU losses from removing captions ($p < 0.001$, App. H.10), and the exact collapse of TRAJEX to ARI 0.000 under the year controls, which is a deterministic outcome rather than a test. Any standard correction leaves these intact. The second class is the period-text differences of 0.006 to 0.064, which we report for completeness and do not treat as established; App. H.3 shows independently that differences of this size lie at the resolution limit of the metrics themselves. Where an argument depends on a small effect it depends on a ratio rather than on a threshold: the adjacent-period exchange costs 0.059 ARI against 0.445 for a random exchange, and that order-of-magnitude gap is what the argument uses.

H.2 DOES MATCHED-PAIR AVERAGING FLATTER OMITTED PERIODS?

The main evaluation averages temporal IoU over matched pairs, so a method that returns fewer periods is scored on fewer, possibly easier, comparisons. Table 19 repeats the calculation with every unmatched period charged the metric's minimum, using Eq. (2) with $q_{\min} = 0$.

Table 19: Temporal IoU when unmatched periods are excluded from the average (the main protocol) or assigned zero.

	OPEN-MLLM		TRAJEX	
Aggregation	**GT**	**Pred.**	**GT**	**Pred.**
Matched periods only	0.464	0.352	0.633	0.394
Unmatched periods assigned zero	0.386	0.278	0.542	0.258

All four scores fall, and the largest drop is TRAJEX with Pred. input, from 0.394 to 0.258, consistent with its heavy under-segmentation there. The choice of aggregation is not neutral: under the stricter variant OPEN-MLLM becomes the higher-scoring method with Pred. input (0.278 against 0.258), reversing the ordering in Tab. 4. We keep the matched-pair average in the main results because it isolates range accuracy given a correspondence, and because grouping, count, and sequence metrics already charge for missing and additional periods (Tab. 12). The two aggregations answer different questions, and we report the reversal because the choice between them changes which baseline leads.

H.3 HOW MUCH CAN THE TEXT METRICS RESOLVE?

BERTScore and MoverScore are reported for six free-text outputs, so it matters what a given difference between two scores can mean. We compare each generated text with four references: its matched reference, another period of the same topic where one exists, a period of an unrelated topic, and itself. To place the two metrics on a common scale we normalise,

$$\tilde{q} = \frac{q - \bar{q}_{\text{unrelated}}}{\bar{q}_{\text{self}} - \bar{q}_{\text{unrelated}}}, \tag{5}$$

so an unrelated topic scores 0 on average and an exact copy 1. These values are diagnostic only; Tab. 4 reports raw scores. We also report the area under the ROC curve for separating the matched reference from an alternative, where 0.5 is chance and 1.0 perfect separation.

BERTScore has a high floor. Text from an unrelated topic already scores 0.826 to 0.838. Every BERTScore in Tab. 4 lies between 0.83 and 0.89, so almost all of the raw value is floor and only the last few hundredths carry information. MoverScore floors are near zero (-0.009 to 0.060), so

Table 20: What BERTScore (BS) and MoverScore (MS) can resolve. Top: the raw score already obtained against an unrelated topic, the normalised score of the matched reference from Eq. (5), and the pair-level Spearman correlation ρ between the two metrics. Bottom: separation as AUC. "–" marks a within-topic comparison that does not apply, because topic-level and caption outputs have no sibling period.

Output	Unrelated-topic score BS / MS	Normalised matched score BS / MS	BS–MS agreement ρ
Caption	0.838 / 0.060	0.105 / 0.080	0.75
Period title	0.837 / −0.008	0.050 / 0.048	0.46
Period description	0.835 / 0.032	0.101 / 0.093	0.80
Topic title	0.826 / −0.002	0.306 / 0.293	0.73
Topic description	0.828 / −0.009	0.062 / 0.049	0.70

Output	Matched vs. unrelated topic AUC BS / MS	Matched vs. other period AUC BS / MS
Caption	0.88 / 0.93	–
Period title	0.58 / 0.63	0.54 / 0.58
Period description	0.84 / 0.92	0.61 / 0.64
Topic title	0.85 / 0.93	–
Topic description	0.70 / 0.81	–

its small-looking values, such as 0.35 for a topic title, occupy much more of the usable range. After Eq. (5) removes the floor, the two metrics separate the alternatives by similar amounts, which is why we report both.

Separating periods is much harder than separating topics. For period descriptions, AUC is 0.84/0.92 against an unrelated topic but only 0.61/0.64 against another period of the same topic. In raw terms, MoverScore averages 0.123 for the matched period description, 0.098 for a different period of the same topic, and 0.032 for an unrelated topic. The gap Task 2 must resolve is the 0.025 between the first two, not the 0.091 to the third, and App. H.9 uses that 0.025 as its reference scale.

Short titles carry the weakest signal. Period titles are a few words long. Their AUC is 0.58/0.63 against an unrelated topic and 0.54/0.58 against a sibling period, barely above chance, and their BS–MS correlation is the lowest at $\rho = 0.46$. Differences in period-title scores in Tab. 4, such as the 0.058 MoverScore gap favouring OPEN-MLLM, therefore do not on their own indicate better trajectories. This is the concrete basis for the text-metric limitation in Sec. 7.

H.4 WHICH DIFFERENCES ARE CONSISTENT ACROSS TOPICS?

Table 21: Task 1 means with 95% confidence intervals across topics. The final column names a baseline only where its paired advantage is consistent across topics ($p < 0.05$); "–" means neither baseline has one. Macro-F1 and caption metrics are omitted because the main runs did not retain the image-level scores this analysis needs.

Metric	Open-MLLM	TrajEx	Consistent advantage
Year MAE ↓	4.72 [4.29, 5.15]	4.65 [4.27, 5.05]	–
Environment accuracy ↑	0.257 [0.227, 0.288]	0.265 [0.242, 0.291]	–
Activity accuracy ↑	0.279 [0.260, 0.299]	0.315 [0.293, 0.335]	TrajEx (0.036)

On Task 1 the baselines are distinguishable only on activity accuracy, where TRAJEX leads by 0.036; the year and environment intervals overlap substantially (Tab. 21). The observation in Sec. 6 that TRAJEX is numerically better on every Task 1 output should be read against this: one of those three advantages is consistent across topics.

Table 22 separates two questions that Tab. 4 answers together: which baseline is better under the same input setting, and how much each loses when its own predictions replace ground truth.

With GT input the baselines separate cleanly and in opposite directions: TRAJEX on trajectory structure, OPEN-MLLM on trajectory semantics. With predicted input only period count and period-title MoverScore still differ consistently, and the period-title gap of 0.018 lies inside the range App. H.3

Table 22: Paired Task 2 differences that are consistent across topics. Parentheses give the absolute gap between baselines. *Loss* is the positive magnitude of deterioration when predicted inputs replace GT inputs. "–" marks a difference that is not consistent ($p \geq 0.05$). One metric per structural output is shown; Tab. 4 gives the complete set.

	Consistent baseline advantage		*Loss with predicted input*	
Metric	**GT**	**Pred.**	**Open-MLLM**	**TrajEx**
Image grouping (ARI ↑)	TrajEx (0.232)	–	0.153	0.337
Period count ($\lvert\hat{J} - J\rvert$ ↓)	–	Open-MLLM (1.098)	–	0.854
Period range (temporal IoU ↑)	TrajEx (0.169)	–	0.112	0.239
Full sequence (Lev. ↓)	–	–	0.215	0.210
Period title (MS ↑)	Open-MLLM (0.058)	Open-MLLM (0.018)	0.064	0.024
Period description (MS ↑)	Open-MLLM (0.020)	–	0.033	0.021

shows the metric cannot resolve. The loss columns explain the convergence: TRAJEX loses more than twice as much ARI as OPEN-MLLM (0.337 against 0.153), the gap analysed in App. H.7. Predictions worsen all six reported TRAJEX metrics and five of six for OPEN-MLLM.

H.5 NON-VISUAL RULES AND THE EQUAL-SIZE SHORTCUT

Task 1 reference rules. Table 23 evaluates rules that predict a constant without inspecting the image. The median year is the median over the ground-truth years of all 20,100 images, and the midpoint rule is the centre of the benchmark's 1840–2025 span. The per-topic median is an oracle, since it needs the ground-truth topic of the image it predicts.

Table 23: Task 1 reference rules evaluated on all 20,100 images. "–" marks an output the stated rule does not predict.

Rule	**Year MAE**	**Accuracy / macro-F1**
Constant median year (1988)	31.68	–
Constant midpoint of the benchmark span (1933)	55.10	–
Per-topic median year (oracle)	28.50	–
Majority environment (*portrait studio*)	–	0.013 / 0.000
Majority activity (*Working at a workplace*)	–	0.023 / 0.000

The constant-year rules give MAE between 28.50 and 55.10 years, six to twelve times the roughly 4.7 years of both baselines, and even the per-topic oracle does not approach them. Majority-class prediction is uninformative for environment and activity, at accuracy near 0.02 and macro-F1 of 0.000 under the full 669- and 177-class label spaces.

The equal-size shortcut. Because every period contains exactly 100 images, sorting a topic's images by year and cutting the sorted list into J contiguous bins of equal size reconstructs the ground truth almost perfectly. To size this shortcut and check that it is not what the baselines exploit, we subsample each period to a random fraction between 40% and 100% of its images under a fixed seed, leaving 14,339 of the 20,100 images, and rerun both the shortcut rules and the baselines.

The shortcut is real but is not what the baselines use. With GT years and the true count, equal-count bins reach ARI 0.999 on the full collection and 0.715 once period sizes vary, while the baselines barely move over the same change: OPEN-MLLM from 0.562 to 0.554 and TRAJEX from 0.794 to 0.758. On the unequal-size collection TRAJEX overtakes even the oracle equal-count rule; OPEN-MLLM does not.

Three points follow, and they are why this diagnostic is in the paper.

Equal period size is a construction artifact, not a property of the task. Every period holds 100 images regardless of its duration, so period mass is unrelated to period span; among the example topics of App. G.3 alone, spans run from six years to eighty-five. That single choice is the largest determinant of achievable ARI: it is what moves equal-count binning from 0.715 to 0.999. Grouping metrics should therefore be reported on the unequal-size collection as well as the equal-size one, since a score on the equal-size collection alone cannot be separated from the shortcut; the subsample is determined by the seed and per-period fractions given above. Sampling images in proportion to

Table 24: Task 2 ARI before and after removing the equal-size regularity. The equal-count rules sort a topic's images by supplied year and cut the sorted list into contiguous bins of equal size. Rules given the true J are oracles; the $J = 5$ rule uses the collection median and no topic-specific ground truth.

Method or rule	Supplied years	100 per period	Unequal size
Equal-count bins, true J (oracle)	GT	0.999	0.715
Equal-count bins, $J = 5$	GT	0.937	0.705
Equal-count bins, true J (oracle)	Open-MLLM Pred.	0.742	0.660
Equal-count bins, $J = 5$	Open-MLLM Pred.	0.713	0.653
Equal-count bins, true J (oracle)	TrajEx Pred.	0.756	0.631
Equal-count bins, $J = 5$	TrajEx Pred.	0.723	0.625
Open-MLLM	GT	0.562	0.554
TrajEx	GT	0.794	0.758

period duration would remove the artifact at its root, but requires regeneration rather than resampling and is left to the next release.

Equal-count binning bounds one metric; it is not a method. It consumes the Task 1 years, so it is not year-free: with GT years it is an oracle, and with predicted years it inherits a vision model's dating. It returns image assignments, the ranges derived from them, and their order, which is three of Task 2's seven outputs and none of the four free-text ones. It identifies no topic and writes no period title or description.

Given accurate dates, Task 2 grouping is a segmentation problem. With predicted years and no topic-specific ground truth it reaches 0.625 to 0.660, above the end-to-end ARI of both baselines (0.41 and 0.46 in Tab. 4). On that metric a fixed rule over predicted years is stronger than either learned method, and we report it because locating the difficulty is what the diagnostic is for: on VDBench the hard part of Task 2 grouping is year estimation and boundary placement rather than the grouping model, so a method earns its place on the outputs where no such rule applies. App. H.10 gives the evidence on that side, where removing captions collapses OPEN-MLLM's period-description MoverScore from 0.144 to 0.034 against an unrelated-reference floor of 0.032.

H.6 TASK 1 ROBUSTNESS

Year error by era. Define signed error as predicted minus ground-truth year, so positive means dated too late. Across all 20,100 images both baselines are worst before 1900, at MAE 15.58 years for OPEN-MLLM and 12.32 for TRAJEX; the mean signed errors there are $+14.21$ and $+8.03$ years, so much of the error is systematic bias rather than scatter, pronouncedly so for OPEN-MLLM (14.21 of 15.58 years). For 1946–1979 the signed errors are -6.04 and -3.36, meaning dates that are too early. Errors are lower after 1980. Because 42% of the collection dates from 2000 onward, the overall MAE in Tab. 4 is weighted towards recent images and understates the difficulty of the nineteenth-century material.

Image perturbations. To test whether dating rests on a single image cue, we rerun Task 1 on the fixed 1,025-image subset under three transformations: greyscale conversion, which removes colour; low-pass filtering by downsampling eightfold and upsampling again, which removes fine texture; and random permutation of a 4×4 patch grid, which destroys global layout while preserving local content. Perturbed images are stored as JPEG at quality 95, near the least-compressed end of the 0–100 scale, while the originals keep their native files, so a small part of each measured change is re-encoding rather than the perturbation. OPEN-MLLM reruns its complete Task 1 procedure; TRAJEX recomputes year, environment, and activity but does not regenerate captions.

On this subset the unperturbed year MAE is 4.52 for OPEN-MLLM and 4.62 for TRAJEX. The two degrade differently: low-pass filtering is worst for OPEN-MLLM, adding 2.20 years, while patch shuffling is worst for TRAJEX, adding 1.43 years, consistent with a language-model reader using fine detail and a CLIP-style encoder using global layout. Patch shuffling costs both 9.1 percentage points of activity accuracy. The worst perturbed MAE is therefore below 6.8 years, against 31.68 years for the constant median-year rule (Tab. 23): no single colour, texture, or layout cue accounts for the dating accuracy on its own.

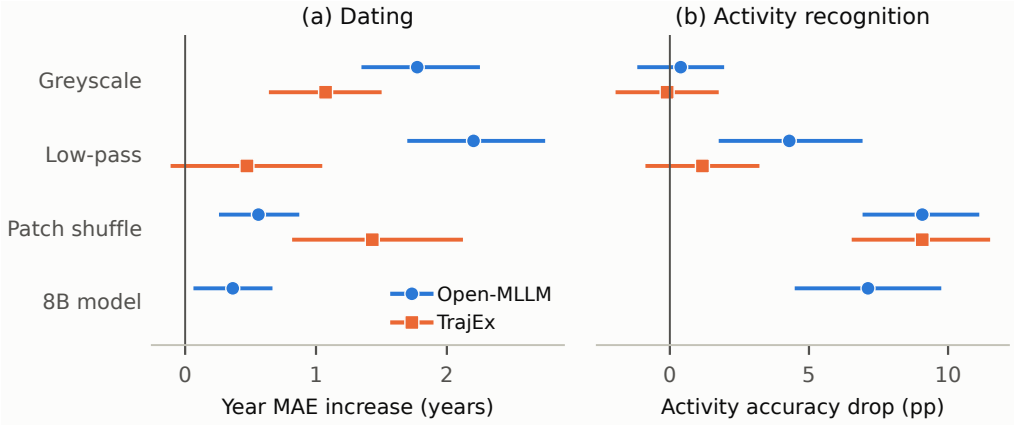

Figure 9: Task 1 change under image perturbations and under OPEN-MLLM's 8B variant, measured on the fixed 1,025-image subset relative to the original images. Panel (a) shows the increase in year MAE, panel (b) the drop in activity accuracy, so a rightward shift is always worse and zero is no change. Error bars are the 95% confidence intervals of App. H.1. The 8B row replaces Qwen3-VL-32B with Qwen3-VL-8B in OPEN-MLLM; TRAJEX has no 8B variant.

Model size. Replacing Qwen3-VL-32B with Qwen3-VL-8B in OPEN-MLLM increases year MAE by 0.36 years and reduces activity and environment accuracy by 7.1 and 6.0 percentage points. In relative terms that is roughly 8% on dating against 23 to 25% on the labels, so scale matters much more for the closed-vocabulary predictions.

H.7 WHICH TASK 2 INPUTS MATTER

We vary whether three Task 1 input groups come from ground truth or from the baseline's predictions — year, the two labels, and caption — and report the main effect of each group, that is, its average effect over the settings of the other two.

Table 25: Reduction in Task 2 image-grouping ARI when one ground-truth Task 1 field is replaced by the baseline's own prediction, averaged over the settings of the remaining fields. Brackets give 95% confidence intervals across topics; "n.s." marks an interval containing zero.

Baseline	Field replaced	Reduction in ARI
TRAJEX	Year t_i	0.337 [0.238, 0.431]
TRAJEX	Environment and activity labels	no measurable change
TRAJEX	Caption c_i	no measurable change
OPEN-MLLM	Caption c_i	0.115 [0.084, 0.148]
OPEN-MLLM	Year t_i	0.028 (n.s.)
OPEN-MLLM	Environment and activity labels	0.017 (n.s.)

The baselines rely on disjoint inputs. TRAJEX depends on the year and on nothing else, which follows from segmenting the year sequence; OPEN-MLLM depends on the caption, and its year and label effects cannot be distinguished from zero. This is also why end-to-end evaluation degrades TRAJEX more (Tab. 22): Task 1 year error is exactly the input it cannot do without.

Controlled year noise. To measure the year dependence directly rather than through coupled prediction error, we add independent Gaussian noise to every ground-truth year for TRAJEX, round to whole years, and clip to the benchmark span. OPEN-MLLM is omitted because the replacement above shows it is insensitive to the year field.

Grouping degrades steeply and monotonically: two years of noise already cost 0.194 ARI. Before rounding and clipping, Gaussian noise with $\sigma = 5$ has an expected absolute error of about 4.0 years, close to TRAJEX's own MAE of 4.65, and the two conditions give ARI 0.490 and 0.457. Real prediction errors are not Gaussian and vary by era (App. H.6) and topic, but the correspondence is close enough to say that TRAJEX's end-to-end score is governed by its dating error. On VDBench, improving year estimation therefore buys more Task 2 grouping accuracy than improving any other Task 1 output.

Table 26: TRAJEX Task 2 image grouping as Gaussian noise is added to the ground-truth years. The final row is the end-to-end setting, where the year error is TRAJEX's own rather than injected.

Supplied years	ARI
Ground truth, no error	0.794
Ground truth + Gaussian noise, $\sigma = 2$ years	0.600
Ground truth + Gaussian noise, $\sigma = 5$ years	0.490
Ground truth + Gaussian noise, $\sigma = 10$ years	0.359
TRAJEX Task 1 predictions, year MAE 4.65	0.457

H.8 TEMPORAL SEGMENTATION VERSUS FEATURE-SPACE CLUSTERING

Since TRAJEX forms periods by segmenting years, the question is whether clustering in a joint visual–textual space would do better, and in particular whether it would recover periods that are visually distinct but temporally adjacent. We replace only the segmentation stage with HDBSCAN (Campello et al., 2013), leaving the rest of the pipeline unchanged. Each image is represented by its DINOv3 visual features, its caption and label embeddings, and one year coordinate. HDBSCAN is given no true period count; its minimum cluster size is 2% of a topic's images, with a floor of three. Both variants run on all 41 topics in both settings; the ground-truth mean is 4.90 periods per topic.

Table 27: TRAJEX Task 2 results with its default temporal segmentation or with HDBSCAN feature-space clustering. Mean $\hat{J}$ is the average number of periods predicted per topic, against a ground-truth mean of 4.90.

Input	Period method	ARI	AMI	Temporal IoU	Mean $\hat{J}$
GT	Temporal segmentation	0.794	0.868	0.633	4.37
GT	Feature-space clustering	0.129	0.184	0.322	3.78
Pred.	Temporal segmentation	0.457	0.580	0.394	3.12
Pred.	Feature-space clustering	0.088	0.127	0.293	2.93

Feature-space clustering is worse on every metric, losing 0.665 ARI with GT input and 0.369 with Pred. input. Its failure is under-segmentation: it predicts fewer periods than ground truth for 31 of 41 topics with GT input and 32 of 41 with Pred. input, collapsing a topic to a single period for 13 and 11 respectively, while over-segmentation is rarer. A follow-up diagnostic on eight topics found that DINOv3 features alone do not recover the period structure, and that raising the weight on the year coordinate helps some topics only by moving the representation towards temporal segmentation. TRAJEX therefore keeps temporal segmentation, and current general-purpose visual features do not by themselves expose the period structure VDBench asks methods to recover.

H.9 TIME AND CONTENT CONTROLS

These controls break one input channel at a time to separate what each baseline takes from the year signal from what it takes from image content. In every case the permutation is within a topic and the ground-truth period of each image is preserved, so only the method's information is degraded, never the evaluation target.

Year controls. Table 29 reports the year controls with GT Task 1 input. Constant year is reported only for TRAJEX, since OPEN-MLLM is insensitive to the year field (Tab. 25).

TRAJEX drops to ARI exactly 0.000 under both year controls, the value expected of a grouping unrelated to the ground truth, confirming it has no non-temporal grouping mechanism. OPEN-MLLM retains ARI 0.176 after shuffling, about a third of its GT score, because its captions and labels still carry period information once the years are meaningless. Both retain most of their period-description MoverScore, which reflects the narrow usable range of that metric (App. H.3) rather than preserved trajectory quality.

Content controls. Random shuffle is evaluated with GT and Pred. inputs; adjacent-period and pixel shuffles use GT input only. Pixel shuffle applies to TRAJEX alone, since OPEN-MLLM receives no pixels in Task 2.

Table 28: The five controls. All permutations are within a topic, and the ground-truth period of each image is always preserved.

Control	What is altered	What it tests
Year shuffle	The supplied years are permuted across the topic's images; each image keeps its own pixels, caption, and labels	Whether grouping is driven by the year field alone
Constant year	Every image is given the same year	What remains once within-topic time information is removed entirely
Random content shuffle	Pixels, caption, and labels are permuted across the topic's images; each image keeps its own year	Whether content beyond the year carries period information
Adjacent-period shuffle	As above, but replacement content is drawn only from a neighbouring period	Whether that content signal is fine-grained or only separates broad eras
Pixel shuffle	Only the pixels are permuted; year, caption, and labels stay with the image	Whether TRAJEX uses visual features for grouping at all

Table 29: Task 2 results after shuffling or removing the supplied year signal, with GT Task 1 input. Lev. is the normalised Levenshtein distance and Descr. MS the period-description MoverScore.

Baseline	Condition	ARI	AMI	$\lvert\hat{J} - J\rvert$	Temporal IoU	Lev.	Descr. MS
Open-MLLM	GT reference	0.562	0.661	0.634	0.464	0.236	0.144
Open-MLLM	Shuffled years	0.176	0.206	0.780	0.220	0.411	0.133
TrajEx	GT reference	0.794	0.868	1.171	0.633	0.196	0.124
TrajEx	Shuffled years	0.000	0.000	1.098	0.143	0.697	0.108
TrajEx	Constant year	0.000	0.000	3.902	0.000	0.795	0.101

Table 30: Task 2 grouping and period-description results under the content controls. Deltas are relative to the matching reference row. Bold marks a paired change that is consistent across topics ($p < 0.05$). "–" in a reference row means no change from itself is reported.

Baseline and input	Condition	ARI	Δ ARI	Descr. MS	Δ MS
Open-MLLM GT	Reference	0.562	–	0.144	–
Open-MLLM GT	Random content shuffle	0.117	**−0.445**	0.117	**−0.027**
Open-MLLM GT	Adjacent-period shuffle	0.503	**−0.059**	0.124	**−0.020**
Open-MLLM Pred.	Reference	0.409	–	0.111	–
Open-MLLM Pred.	Random content shuffle	0.282	**−0.127**	0.097	**−0.014**
TrajEx GT	Reference	0.794	–	0.124	–
TrajEx GT	Random content shuffle	0.774	−0.020	0.109	**−0.015**
TrajEx GT	Pixel shuffle	0.794	+0.000	0.117	**−0.007**
TrajEx GT	Adjacent-period shuffle	0.747	−0.047	0.111	**−0.013**
TrajEx Pred.	Reference	0.457	–	0.103	–
TrajEx Pred.	Random content shuffle	0.507	+0.050	0.097	**−0.006**

OPEN-MLLM does use content for grouping: random shuffling costs it 0.445 ARI with GT input and 0.127 with Pred. input. That content signal is coarse, however. Restricting the exchange to neighbouring periods reduces the GT effect from 0.445 to 0.059, so most of what the content contributes is the separation of broad eras rather than the placement of a boundary between adjacent periods, which is the distinction Task 2 actually requires. TRAJEX grouping shows no consistent change under any content control, including pixel shuffle at exactly ± 0.000, independently confirming Tab. 25: its image features influence only period merging and medoid selection, never the assignment. TRAJEX Task 2 grouping is therefore a one-dimensional segmentation of the year sequence, and its visual components act on the semantic outputs rather than on which image goes where.

Period-description MoverScore falls under every content control, so both baselines do use content when writing descriptions. The margins are small in absolute terms, but the largest, 0.027 for OPEN-MLLM with GT input, is comparable to the 0.025 separating a matched period description from a wrong-period one (App. H.3). Relative to what the metric can resolve, the drop is close to the maximum observable.

H.10 CAPTION CONTENT AND EXPLICIT DATES

Table 25 shows OPEN-MLLM depends on the caption, which invites an objection: 5,025 of the 20,100 ground-truth captions (25.0%) state a year or decade in words, so the model might simply be reading dates out of text. This control removes either the date expressions or the whole caption.

Table 31: OPEN-MLLM Task 2 results with GT input after removing dates from, or removing entirely, the ground-truth captions.

Caption input	ARI	Temporal IoU	Period title MS	Period descr. MS
Original	0.562	0.464	0.088	0.144
Explicit dates removed	0.532	0.475	0.077	0.142
No caption	0.404	0.360	−0.007	0.034

Removing the dates produces no change that is consistent across topics, whereas removing the captions costs 0.158 ARI and 0.104 temporal IoU ($p < 0.001$ for both) and collapses the period texts, with period-title MoverScore falling to -0.007, essentially the -0.008 it scores against an unrelated reference (Tab. 20). OPEN-MLLM reads the described visual context, not the dates written in it.

The explicit dates remain a limitation of the caption field rather than of the method. Because captions are written alongside the generation prompt (App. A.4), a quarter of them contain dates that no method can recover from pixels, so Task 1 caption scores carry an unrecoverable component. Future releases should generate captions without explicit dates, or mask them before scoring.

H.11 SUPPLYING THE TRUE PERIOD COUNT

App. G.1 attributes TRAJEX's errors mainly to count selection and OPEN-MLLM's mainly to image assignment. Supplying the true count J tests that attribution directly.

Table 32: OPEN-MLLM Task 2 with and without the true period count added to the instruction. "Used supplied count" is the number of topics whose returned count matched the one supplied. "–" means no count was supplied.

Input	Count	Used supplied count	$\lvert\hat{J} - J\rvert$	ARI	Temporal IoU
GT	Unconstrained	–	0.634	0.562	0.464
GT	True J supplied	36 / 41	0.171	0.575	0.491
Pred.	Unconstrained	–	0.927	0.409	0.352
Pred.	True J supplied	40 / 41	0.024	0.406	0.370

OPEN-MLLM follows the supplied count in 36 of 41 topics with GT input and 40 of 41 with Pred. input, so count error nearly vanishes, yet ARI moves by only $+0.013$ and -0.003. Knowing how many periods exist does not tell it which images belong to which. TRAJEX responds in the opposite way: given the true count it reaches ARI 0.968 with GT years against 0.794 unconstrained,

and 0.712 with predicted years against 0.457. Almost all of its remaining GT-input error is count selection, while with predicted years the residual gap is inherited year error (Tab. 26).

The two failure modes are complementary and neither is fixed by the other's remedy. A method combining TRAJEX's temporal precision with OPEN-MLLM's ability to choose a plausible number of periods would improve on both, and none is provided here.

I EXTENDED RELATED WORK

I.1 DIACHRONIC ANALYSIS IN LINGUISTICS

Diachronic linguistics examines how language changes over time (Saussure, 1959). Historical semantics focuses on changes in word meaning, including the acquisition or loss of senses by the same lexical form (Blank, 1999; Traugott & Dasher, 2002). Computational methods for lexical semantic change compare uses of a target word across dated text collections, often through representations learned separately for each period (Kutuzov et al., 2018; Tahmasebi et al., 2021). Hamilton et al. (2016), for example, align historical word embeddings and trace changes in nearest-neighbour structure across decades. SemEval-2020 Task 1 established a shared evaluation in four languages: systems determined whether a word had changed meaning between two periods and ranked target words by degree of change (Schlechtweg et al., 2020). This is the conceptual and evaluation precedent for visual diachrony, but it concerns words in text rather than subjects in images. Later SemEval tasks connect lexical meaning with visual evidence. SemEval-2023 Task 1 asks a system to select, from ten candidates, the image depicting the intended sense of an ambiguous word (Raganato et al., 2023); SemEval-2025 Task 1 asks it to rank images by how well they express the idiomatic or literal reading of an expression (Pickard et al., 2025). Neither models change over time.

I.2 VISUAL CONTEXT AND COMPOSITIONAL UNDERSTANDING

Visual context has long supported recognition (Galleguillos & Belongie, 2010; Wang & Zhu, 2023). Scene recognition identifies the type of place shown, such as a kitchen or a street (Zhou et al., 2018); its Places365 labels also contribute to the VDBench environment vocabulary (App. B.1). Scene-graph generation predicts relations between depicted entities (Li et al., 2024), and situation recognition predicts an activity together with participant roles (Yatskar et al., 2016). Yuksekgonul et al. (2023) show that vision-language models can behave like bags of words and miss compositional relations, and MetaShift organises image subsets by changing object–context correlations to evaluate contextual distribution shift (Liang & Zou, 2022). These works address components of context, or robustness to changes in it, but not the historical development of one subject's context and meaning.

I.3 DATING HISTORICAL IMAGES

A first strand of temporal computer vision treats appearance as evidence for an image's date. Palermo et al. (2012) predict a photograph's decade from properties of its colour imaging process. Ginosar et al. (2017) estimate the capture year of yearbook portraits, and Salem et al. (2016) date photographs of people from clothing, hairstyles, and glasses. Vittayakorn et al. (2017) estimate when depicted cars and garments were manufactured. These methods infer time from appearance but do not recover a sequence of historically interpreted contexts.

I.4 MINING VISUAL CHANGE

A second strand traces change across collections. Schindler et al. (2007) recover the relative order of photographs from reconstructed 3D structure. Lee et al. (2013) discover recurring visual elements that vary across time and location, and StreetStyle measures changes in clothing attributes across cities and years (Matzen et al., 2017). Visual Chronicles uses multimodal LLMs to identify frequent co-occurring changes in street-view images collected over a decade (Deng et al., 2025). Their outputs are orderings, changing attributes, or collection-level trends rather than source-grounded period descriptions of a subject's changing meaning.

I.5 Temporal Representations and Distribution Shift

A third strand studies representations and model behaviour over time. Diachronic cross-modal embeddings organise image–text instances while preserving temporal alignment (Semedo & Magalhães, 2019). Barancová et al. (2023) evaluate OpenCLIP representations for dating historical press photographs. Tekaya et al. (2025) show that time forms a low-dimensional, non-linear structure in vision-language representations from which image year can be estimated; TrajEx uses their temporal projection (App. F). CLEAR evaluates classification as visual concepts evolve (Lin et al., 2021), Wild-Time studies temporal distribution shift in five real-world application datasets (Yao et al., 2022), and TiC-CLIP benchmarks continual training on web image–text pairs dated from 2014 to 2022 (Garg et al., 2024). These benchmarks test temporal encoding or robustness, not period-by-period discovery of visual and semantic context.

I.6 Synthetic Data and Historical Image Generation

Synthetic images are widely used when real images or dense annotations are difficult to obtain. For Kinect pose estimation, Shotton et al. (2011) use varied synthetic depth images to train body-joint prediction. He et al. (2023) train image classifiers on text-to-image outputs when real data is scarce, and models for face recognition (Kim et al., 2023) and 3D body estimation (Black et al., 2023) have been trained entirely on synthetic humans. In these settings, synthesis provides scale and task-specific labels.

Historical generation adds a difficulty: an image can look convincing while containing details that are wrong for its stated period. Existing methods address narrower transformations. Wan et al. (2020) restore degradation in old photographs, Time-Travel Rephotography reconstructs an antique portrait as if a modern camera had captured it (Luo et al., 2021), and Chen et al. (2023) resynthesise portraits across decades while preserving identity. All three start from a supplied photograph and do not construct varied scenes for a source-grounded trajectory.

General text-to-image models introduce further risks: they can amplify demographic stereotypes (Bianchi et al., 2023), distort societal representation (Luccioni et al., 2023), and show substantial cross-country gaps in cultural awareness (Kannen et al., 2024). Synthetic History evaluates generated depictions of the past directly and identifies era-specific default styles, anachronisms, and implausible demographic distributions (De Rosa Palmini & Cetinic, 2026). These findings are what make the grounding and validation steps of Apps. B and D necessary when generated images serve as benchmark data.

I.7 Position of VDBench

To the best of our knowledge, no existing benchmark combines source-grounded historical periods with per-image context labels to evaluate both individual depictions and complete diachronic trajectories. VDBench sits at the intersection of the strands above: its research question comes from lexical semantic change detection, its per-image temporal labels from image dating, its context vocabularies from scene and activity recognition, and its images from controlled generation, and it adds the trajectory-level evaluation that none of them provides.