

MEMESOVERTIME: A Decade-Scale Study of Meme Dynamics

Shijia Zhou[▲][✉]

Diego Frassinelli^{▲*}

Barbara Plank[▲]^{✉*}

[▲] MaiNLP, Center for Information and Language Processing, LMU Munich, Germany

[✉] Munich Center for Machine Learning (MCML), Munich, Germany

{zhou.shijia, diego.frassinelli, b.plank}@lmu.de

Abstract

Internet memes evolve as they are repurposed across online communities, yet studying how their usage and language shift requires corpora that follow the same templates over many years, which existing resources rarely provide. To address this gap, we collect a decade of posts from Reddit meme subreddits (2016–2025) and introduce MEMESOVERTIME, a semi-automatically curated dataset of 6.6M memes annotated with 6,584 unique templates. We use it to conduct an empirical study of meme template dynamics. We find that templates differ markedly in their temporal activity patterns: most reach peak usage within the first third of their active period, with short-lived templates peaking especially early. For long-lived templates, lexical drift is generally modest, with templates with more instances showing greater stability. Case studies reveal that creators spontaneously develop conventionalized expressions for some templates, while topical references shift over time.

1 Introduction

Internet memes are, by definition, evolving entities. The term *meme* itself, coined by Dawkins (1976) in analogy to the biological gene, implies a unit of cultural information that replicates, mutates, and competes for survival. In online communities, this evolution is not passive: users actively remix, re-contextualize, and repurpose meme *templates* at massive scale, producing millions of *instances* that vary in text, image, and intent while remaining recognizable as belonging to the same template.

Like the Ship of Theseus, “image macros vary along a typicality cline with regard to the degree of multimodality, multilingualism, intra-genre intertextuality and external referencing” (Zenner and Geeraerts, 2018). As Figure 1 illustrates, whether the overlaid content is text alone (top row) or a



Figure 1: In this work, we define a meme template not by visual or textual similarity alone, but by the interplay of its multimodal elements. The six memes above differ markedly in text, images, and topics, while all instantiate the same construction: X exposes Y as Z .

combination of text and image (bottom row), the underlying construction X exposes Y as Z remains constant. In the center of the bottom row, Z is instantiated not by text but by a pineapple image, playing on the fact that “Bananas” without the “B” yields “ananas”, the word for pineapple in many languages. Even within a single template, textual and visual variation can be remarkably rich, spanning different topics, intentions, and the stylistic conventions that individual communities develop over time. This makes determining the boundaries of a meme template operationally challenging.

However, any longitudinal study of meme evolution must first fix a stable unit of analysis, a challenge that, to the best of our knowledge, no large-scale study has systematically addressed. To fill this gap, we define a *meme template* as an equivalence class of base images onto which user-added text is overlaid to produce meme instances, where two base images are equivalent if the text of one can be overlaid onto the corresponding elements of the

* Equal senior contribution.

other with its meaning fully preserved. Following this definition, we derive templates at scale with a semi-automatic pipeline: we first apply embedding-based clustering to group visually similar images, then employ human annotators to verify and merge clusters that share the same underlying template.

Despite growing interest in memes as linguistic and cultural objects (Dancygier and Vandelanotte, 2017; Bülow et al., 2018; Shifman, 2014), existing studies remain limited in scale. Corpus sizes rarely exceed a few thousand instances, and studies of how meme templates transform and adapt over time are virtually absent (see §2.2). This leaves three fundamental questions unanswered:

- **RQ1:** What are the activity dynamics of meme templates over time?
- **RQ2:** What characterizes templates that attract high usage volume compared to those that remain marginal?
- **RQ3:** To what extent does the lexicon of a meme template shift over time?

To answer these research questions, we introduce **MEMESOVERTIME**,¹ the largest meme template corpus to date: 6.6M English-language meme instances from Reddit (2016–2025), organized into 6,584 unique templates. Building on this resource, we conduct a longitudinal study of meme template activity and lexical change. Our contributions are as follows:

- **Dataset and Construction Method:** We introduce MEMESOVERTIME, the largest meme template corpus to date, along with a semi-automatic, data-driven pipeline for template curation combining embedding-based clustering with human verification (§3).
- **Popularity Dynamics:** We provide a large-scale characterization of meme template activity over time, uncovering diverse temporal patterns including viral bursts, seasonal spikes, and sustained usage (§5.1).
- **Lexical Dynamics:** We show that lexical stability scales with template size and longevity, and that drift is driven primarily by turnover in culturally salient references (§5.2).

Taken together, MEMESOVERTIME provides both a large-scale resource and empirical grounding for future work on meme temporal dynamics, template identification, and the evolution of meme templates in online communities.

¹Our code and data are available at:
<https://github.com/mainlp/MemesOverTime>

2 Background

2.1 Temporal studies of internet memes

Internet memes—defined by Shifman (2014) as “(a) a group of digital items sharing common characteristics of content, form, and/or stance; (b) created with awareness of each other; and (c) circulated, imitated, and/or transformed via the Internet by many users”—have become a pervasive mode of digital communication serving as vehicles for humor, political commentary, and collective identity (Vásquez and Aslan, 2021; Schmid et al., 2025; Chagas et al., 2019). The temporal dynamics of memes, including how they spread, persist, and fade, have attracted growing interest in computational research. Existing temporal studies approach this question from three aspects: modeling meme *spread* across networks and platforms (Bauckhage, 2021; Zannettou et al., 2018); predicting *popularity* from content and contextual features (Barnes et al., 2021, 2024); and characterizing *visual patterns* over longer time spans (Lonnberg et al., 2020; Valensise et al., 2021).

Across these lines, however, studies are either limited in scale, restricted to short observation windows, or lack verification of what constitutes a stable unit of analysis over time. MEMESOVERTIME addresses all three: it is larger than prior corpora, spans a full decade, and grounds template identification in a validated annotation pipeline. Building on this, we conduct a large-scale analysis of how meme templates evolve in their temporal frequency, lexical diversity, and visual variation over their lifespan.

2.2 Computational approaches to meme understanding

Earlier computational meme research is largely task-driven (Nguyen and Ng, 2024), encompassing tasks such as sentiment analysis (Sharma et al., 2020; Zhou et al., 2025), hate speech detection (Kiela et al., 2020), offensive content classification (Pramanick et al., 2021), and meme captioning (Hwang and Schwartz, 2023; Zhou et al., 2026). In all these settings, evaluation remains at the instance level, with little attention to the recurring meme templates that gives rise to meaning in the first place.

More recent work recognises the template as the appropriate unit of analysis, introducing it as a proxy for the cultural and social knowledge embedded in memes (Zhou et al., 2024) and for the

Dataset	Templates	Scale
CM50 (Deng et al., 2025)	50	33K
MetaMeme (Lambright et al., 2025)	2,059	274K
IMKG (Tommasini et al., 2023)	2,006	1.3M
SemanticMemes (Zhou et al., 2024)	6,384	3.8M
MEMESOVERTIME (ours)	6,584	6.6M

Table 1: Comparison of meme datasets about templates. **Templates** is the number of distinct templates. **Scale** is the total number of instances.

structured image-text interactions that underpin their meaning (Lambright et al., 2025; Deng et al., 2025). However, prior work adopts two markedly different operational definitions of a template, each carrying its own limitations. CM50 (Deng et al., 2025), MetaMeme (Lambright et al., 2025) and IMKG (Tommasini et al., 2023) derive templates from user-curated websites such as Imgflip,² Know Your Meme,³ and Wikipedia.⁴ This approach inherits the inconsistent naming conventions of these platforms, where the same template may appear under multiple aliases, and it fails to capture memes created outside these platforms, as well as suffering from inevitable coverage lag. SemanticMemes (Zhou et al., 2024) by contrast, derives templates by thresholding perceptual hash similarity, which risk missing instances that have undergone heavy visual editing, such as recoloring, cropping, or overlaid graphics, even when their semantic intent is identical.

To address these limitations, we identify templates through a data-driven clustering pipeline followed by human annotation, without relying on any external catalogue and with greater tolerance for visually diverse instances. As Table 1 shows, MEMESOVERTIME is the largest resource of its kind to date: 6,584 templates over 6.6M instances.

3 Curating MEMESOVERTIME: from Raw Images to Meme Templates

We scrape image posts from Reddit (Baumgartner et al., 2020),⁵ specifically from subreddits containing “meme” in their names. Since the most prominent meme subreddits (e.g., r/meme and r/memes) are English-speaking communities, this study focuses on English-language memes to ensure consistency in temporal analysis. Non-English subred-

dits were filtered out based on their community descriptions and post titles. We collect static images posted on Reddit from January 2016 to December 2025, excluding earlier data due to sparsity.⁶ We retrieve a total of 19M images. Each image is uniquely identified by its Reddit post ID; near-duplicate posts are retained, as repeated usage of the same image constitutes meaningful signal for our temporal analyses.

To identify shared templates across such a large collection, we propose a coarse-to-fine clustering strategy illustrated in Figure 2: a first pass groups memes into broad clusters, and a second pass further partitions each cluster into visually coherent subclusters. Human annotators then consolidate and merge the resulting subclusters to produce the final template labels. We describe the automatic clustering pipeline in §3.1 and the human annotation process in §3.2.

3.1 Clustering and subclustering

3.1.1 Image preprocessing

Since overlaid text is specific to each meme instance rather than the template, retaining it would obscure the underlying template structure during clustering. Therefore, we preprocess all images by masking text and removing margins prior to clustering. Details are provided in Appendix A.1.

3.1.2 Clustering

Following Murgás et al. (2024), who show that CLIP (Radford et al., 2021) combined with UMAP (McInnes et al., 2020) dimensionality reduction and HDBSCAN (McInnes et al., 2017) clustering achieves the strongest performance among unsupervised methods on real-world meme data, we adopt this pipeline for primary clustering. We encode each preprocessed image using CLIP ViT-L/14 to obtain ℓ_2 -normalised embeddings. Unlike k -means, HDBSCAN requires no pre-specified number of clusters, making it well-suited to the long-tailed distribution of meme templates. We apply this pipeline to a month-stratified 5% sample ($\approx 950K$ images) and propagate labels to the remaining instances via k -nearest neighbour voting. This yields 2,406 clusters; full implementation details and hyperparameter selection are provided in Appendix A.2.

⁶Reddit has supported native image hosting only since 2016; earlier image posts were external links, now largely lost to link rot.

²<https://imgflip.com>

³<https://knowyourmeme.com>

⁴<https://www.wikipedia.org>

⁵<https://academictorrents.com/details/3d426c47c767d40f82c7ef0f47c3acacedd2bf44>

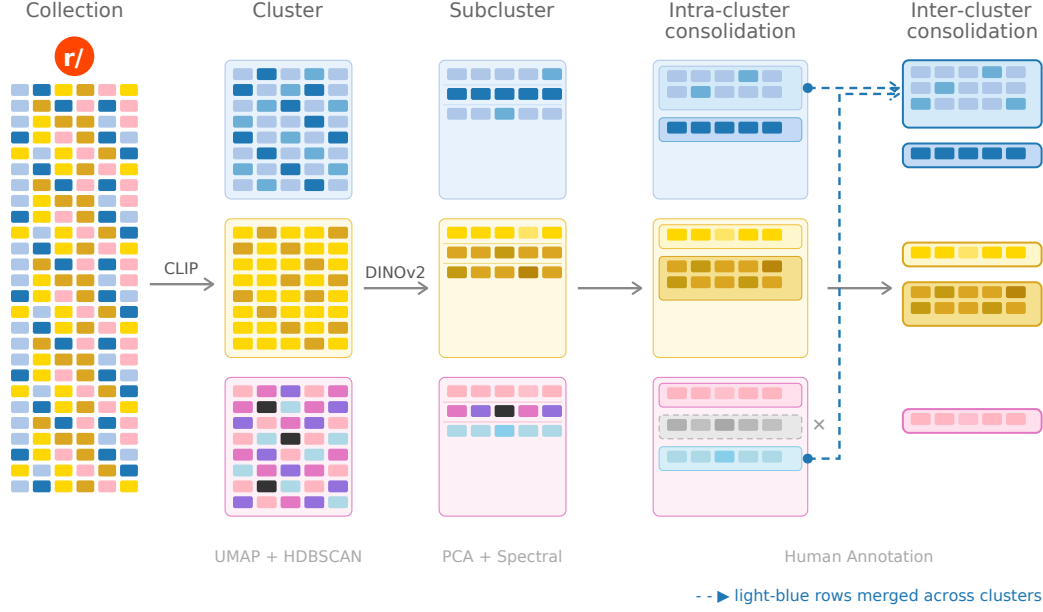


Figure 2: Overview of the MEMESOVERTIME curation pipeline. We collect images, each represented as a single square, from Reddit meme subreddits and apply automatic, data-driven clustering at two levels to group visually similar memes, robust to common template variations such as recoloring, cropping, and overlaid graphics. Human annotators then consolidate and merge clusters to produce the final template labels.

To assess clustering quality, one annotator manually inspected all 2,406 clusters by randomly sampling 100 instances from each and flagging visually inconsistent instances. Of these, 1,184 clusters contained fewer than 5 noisy instances and 683 contained between 5 and 15, indicating that the large majority of clusters are visually coherent and suitable for subsequent annotation.

3.1.3 Subclustering

While primary clustering successfully groups visually similar images, it may still conflate memes whose interpretation-critical elements have been altered. As illustrated in Figure 3, such alterations can fundamentally change a meme’s meaning. To address this, we apply a second stage of subclustering using DINOv2 (ViT-L/14) (Oquab et al., 2024), a purely vision-based model that captures pixel-level visual details, making it sensitive to fine-grained alterations that primary clustering may overlook. We extract ℓ_2 -normalised DINOv2 embeddings, reduce them via PCA retaining 70% of variance, and apply Spectral Clustering (Ng et al., 2001). The optimal number of subclusters $k \in [2, 10]$ is selected by silhouette score. This procedure yields 11,239 visual subclusters across 2,406 primary clusters, with an average of 4.67 subclusters per primary cluster; implementation details are provided in Appendix A.3.

3.2 Consolidating subclusters

Automated clustering captures visual similarity, but visual similarity alone can not determine template identity. We therefore manually annotate each visual cluster to determine whether it constitutes a valid template and whether it should be merged with other clusters sharing the same underlying construction.

We define a meme template as follows: a meme template is a base image (or a class of equivalent base images) onto which user-added text is overlaid to produce meme instances; two base images are considered equivalent if they are identical, or if they are similar and one meme’s user-added text can be overlaid onto the corresponding elements of the other image with the meaning fully preserved.

In practice, this definition translates into two criteria that annotators apply when deciding whether two memes share a template: (i) whether the underlying base image remains consistent across instances, and (ii) whether the meaning is preserved when user-added content is exchanged between instances.

Figure 3 presents representative cases from MEMESOVERTIME where the base image is visually similar but not identical across instances. Under the *Identical* category, such surface-level visual variation leaves the underlying construction intact. Under the *Distinct* category, differences in panel

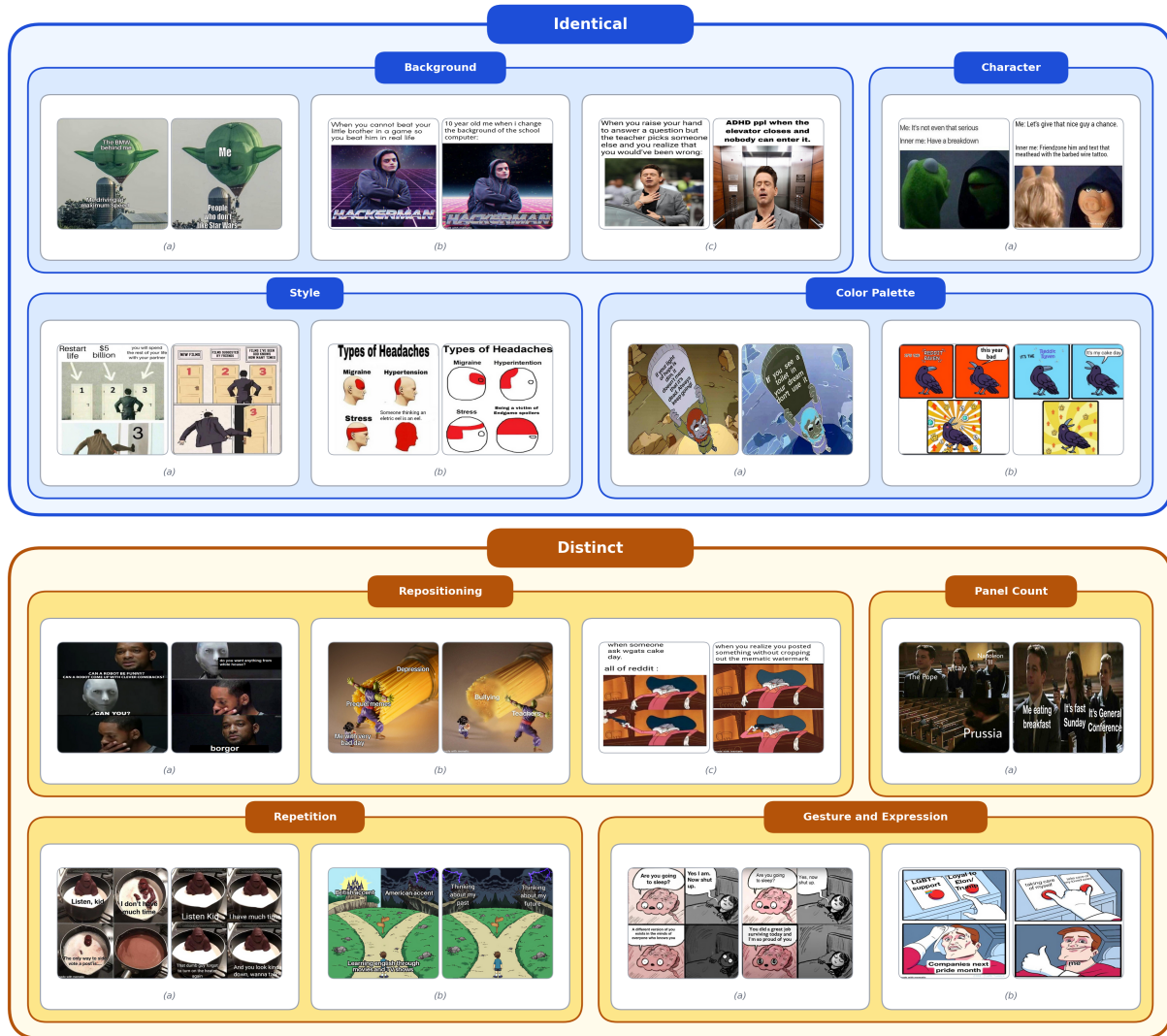


Figure 3: Template boundary decisions in MEMESOVERTIME. Each cell pairs two visually similar memes that are not exact copies, grouped by the aspect in which they differ. Pairs in the top block are annotated as identical templates, those in the bottom block as distinct templates.

structure, object positioning, or gesture are treated as crucial, as they modify the constructional meaning itself. For instance, among the Repositioning examples, reversing the panel order, or removing the protective figure shielding the kid, or redirecting a gun toward oneself each produce a fundamentally different interpretation of the meme. Full annotation guidelines and examples are provided in Appendix A.4. We then conduct an annotation in two stages.

Stage 1: Intra-cluster consolidation For each visual subcluster, we randomly sample 10 instances and exclude the subcluster if (i) the images are not memes (e.g., QR codes, screenshots, game interface captures), or (ii) more than 2 instances are noise. For the remaining subclusters, we annotate which subclusters share the same underlying

template construction and consolidate them into merged clusters.

Stage 2: Inter-cluster consolidation For each merged cluster from Stage 1, we compute DINOv2 embedding centroids and calculate pairwise cosine similarities across clusters. We manually review the 843 pairs where both similarities exceed 0.9, merging those that share the same underlying template construction.

Both annotation steps were carried out by two hired annotators, one of them being an author of the paper. During inter-cluster consolidation, annotators jointly labeled 1,200 instances, achieving a Cohen’s κ of 0.87 (see Appendix A.5 for details). Templates with fewer than 40 instances are then discarded, yielding a final set of 6,584 unique meme templates covering 6,606,128 images.

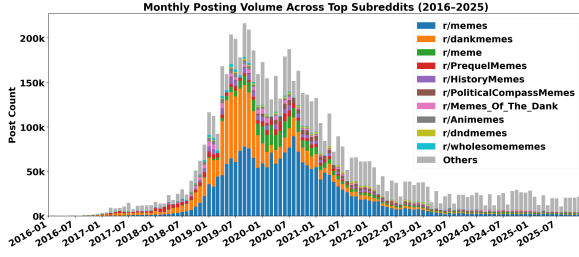


Figure 4: Monthly distribution of all 6.6M instances in MEMESOVERTIME. The ten largest subreddits are shown individually; all remaining subreddits are aggregated into *Others*. Volume is dominated by r/memes and r/dankmemes throughout the decade.

Template size	Count	%	Template size	Count	%
50–100	1,892	30.7	1,000–2,000	556	9.0
100–200	1,230	19.9	2,000–5,000	441	7.1
200–500	1,145	18.6	5,000–10,000	158	2.6
500–1,000	646	10.5	>10,000	103	1.7

Table 2: Distribution of template sizes in MEMESOVERTIME. P25/P50/P75 denote the 25th, 50th (median), and 75th percentiles: P25 = 77, P50 = 170, P75 = 641; mean = 1,003, max = 111,756.

4 The MEMESOVERTIME Dataset

MEMESOVERTIME comprises 6,584 unique meme templates spanning 6.6M images collected from Reddit between 2016 and 2025. Figure 4 shows the monthly posting volume of the full dataset. Volume is dominated by r/memes and r/dankmemes throughout the decade, peaking around 2019–2022. We characterize the dataset along two dimensions: the scale of individual templates (§4.1) and their temporal activity dynamics (§4.2).

4.1 Template size distribution

Table 2 shows the distribution of template sizes in MEMESOVERTIME. The distribution is heavily right-skewed: the majority of templates are small, with 30.7% containing 50–100 instances and a median of 170, while a small number of templates are very large, with 103 templates exceeding 10,000 instances and the largest containing 111,756. The mean size (1,003) is nearly six times the median, reflecting the influence of this long tail.

4.2 Template activity dynamics

We characterize each template along two temporal dimensions: lifespan and active months. **Lifespan** measures the temporal span of a template, defined as the number of months between its first and last observed instance. **Active months** captures how consistently a template appears over time, defined

as the number of distinct months in which at least one instance is observed. Figure 5 presents the distribution of both metrics across all templates in MEMESOVERTIME. The median lifespan is 88 months, with 75% of templates spanning at least 72 months, suggesting that most templates persist over multiple years. Active months, however, reveals more intermittent behaviour: the median is 43 months, meaning that even long-lived templates are frequently absent for extended periods. The two metrics are strongly correlated ($r = 0.73$); we adopt active months as our primary temporal measure, as it naturally supports segmentation into discrete time windows.

5 Meme Template Temporal Dynamics: Insights from MEMESOVERTIME

Using MEMESOVERTIME, we explore two aspects of meme template evolution: how templates vary in their activity patterns over time, and how the text of instances within a template changes across time.

5.1 What are the activity dynamics of meme templates over time?

Meme templates differ not only in how popular they become, but in how long they remain active and whether their usage is sustained or concentrated in short bursts. To characterize these differences, we classify each template into one of four quadrants based on its active months and size relative to the dataset medians (42 active months, 170 instances), yielding four groups: sustained large ($n = 2,803$), bursty large ($n = 494$), sustained small ($n = 489$), and bursty small ($n = 2,798$). The majority of templates fall into either the sustained large or bursty small quadrant (Figure 6a). Manual inspection of the two smaller quadrants reveals characteristic patterns: bursty large templates typically originate from news events (e.g., the first image of a black hole, Nancy Pelosi tearing up a speech), while sustained small templates tend to reflect recurring seasonal or cyclical themes (e.g., Spooky October, Christmas, the NFC season).

We further examine when the activity of a template peaks within its lifespan. For each template, we map its monthly posting activity onto a normalized timeline spanning from its first to last active month, and define the peak as the point of maximum activity. Figure 6b shows the distribution of peak positions across the four quadrants. Across all quadrant types, templates tend to peak

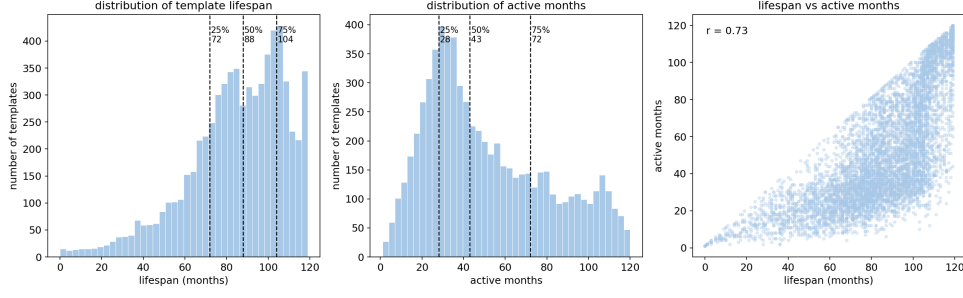


Figure 5: Distribution of template lifespan and active months, and their correlation.

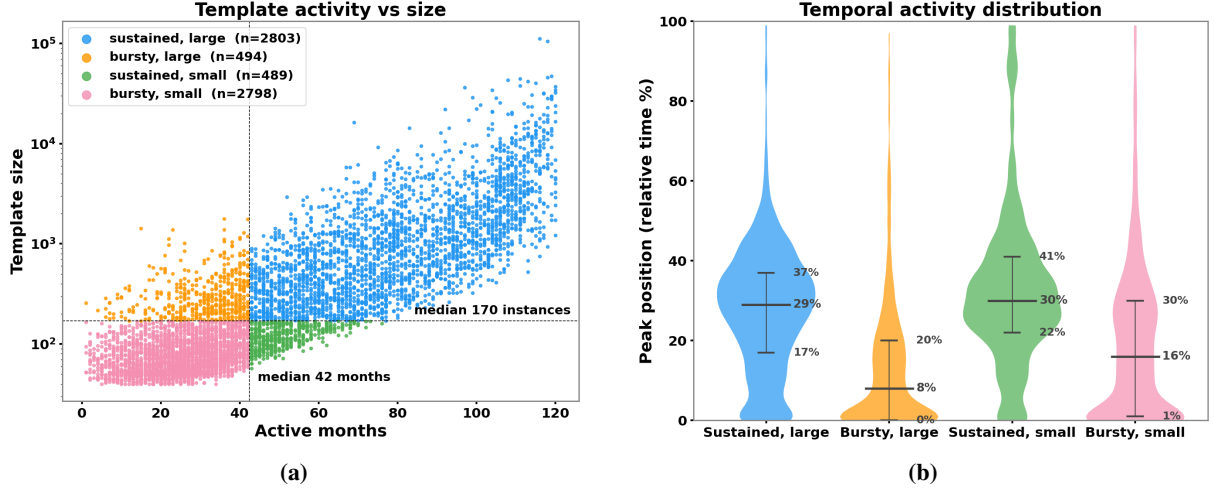


Figure 6: Temporal dynamics of templates in MEMEATLAS. **(a)** Template activity vs size. Each point represents one template, colored by quadrant. Dashed lines mark the median active months and median template size. **(b)** Peak timing distribution across template quadrants. Peak position is measured as the relative position of the month of maximum activity within each template’s active period.

early: the median peak position ranges from 8% to 30% of the active period (this pattern is robust to observation-window censoring; see Appendix B). Bursty templates peak especially early—bursty, large templates show a median of just 8%, indicating that activity is heavily concentrated at onset. Large templates peak earlier than their small counterparts regardless of how long they remain active, suggesting that high-volume templates gain widespread adoption more rapidly.

5.2 To what extent does the lexicon of a meme template shift over time?

Setup For each template, we designate the month with the highest cumulative upvotes as the peak month m^* , and partition its posts into pre-peak, peak, and post-peak segments accordingly. The text content of each post is extracted via OCR using Qwen3-VL (see Appendix A.6 for quality assessment). We restrict our analysis to *sustained-large* templates for which both the pre- and post-peak segments contain at least 500 OCR-extracted posts

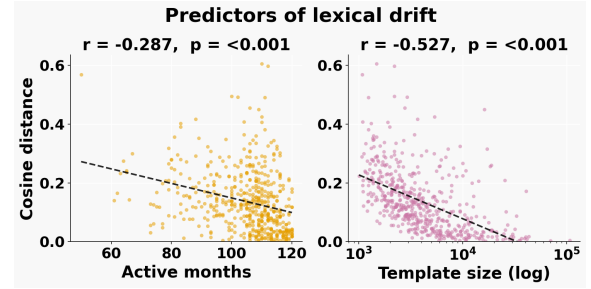


Figure 7: Predictors of lexical drift for sustained-large templates ($n = 531$). Each panel shows the Pearson correlation between cosine distance (pre- vs post-peak) and one template-level feature.

spanning at least 2 active months ($n = 531$).

For each template, we fit a TFIDF model — using sublinear term frequency weighting and unigram–bigram features — on the union of pre- and post-peak text, and measure lexical drift as the cosine distance between the two segment vectors:

$$d = 1 - \frac{\mathbf{v}_{\text{pre}} \cdot \mathbf{v}_{\text{post}}}{\|\mathbf{v}_{\text{pre}}\| \|\mathbf{v}_{\text{post}}\|}. \quad (1)$$

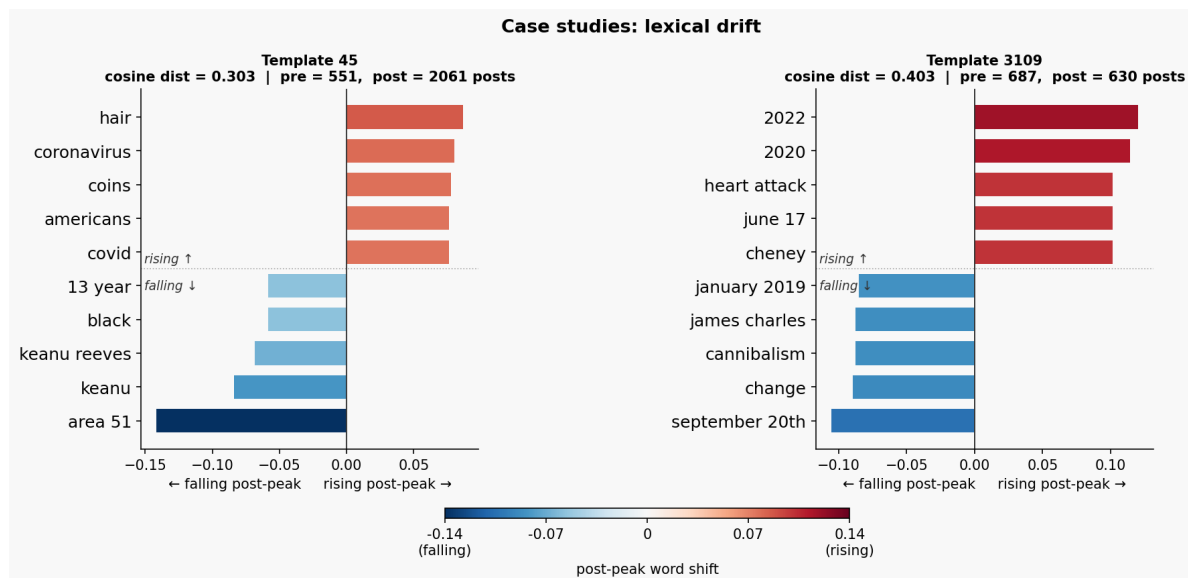


Figure 8: Lexical drift in two sustained-large templates, measured by $\Delta(t) = \text{TFIDF}(t, \text{post}) - \text{TFIDF}(t, \text{pre})$. Positive values indicate terms rising post-peak (red); negative values indicate terms receding (blue).



Figure 9: Instances of Template 45 (a)–(d) and Template 3109 (e)–(h) in MEMESOVERTIME, in chronological order.

The contribution of individual terms is captured by the delta score $\Delta(t) = \text{TFIDF}(t, \text{post}) - \text{TFIDF}(t, \text{pre})$, where positive values indicate terms gaining prominence after the peak and negative values indicate terms receding.

Findings Lexical drift is generally modest (median $d = 0.12$) but exhibits a long tail (max = 0.60), indicating that while most templates retain broadly stable lexicon throughout their lifespan, a subset undergoes substantial change. As shown in Figure 7, both template size and active months correlate negatively with lexical drift. Template size is the stronger predictor ($r = -0.527$): larger templates support a broader vocabulary, which reduces the influence of newly introduced terms on the overall profile. More stable TFIDF estimates for larger templates may also contribute. Active months show a weaker association ($r = -0.287$).

Among templates with above-average drift, the terms that decline most consistently are temporally anchored cultural references. The two case studies in Figure 8 show this pattern concretely: the terms that decline post-peak reflect trending topics that were prominent at the time but have since faded from online discourse. Both templates examined

here fall in the large, sustained quadrant (see §5.1). For Template 45 in MEMESOVERTIME, *area 51* and *keanu reeves* were prominent before the peak yet largely absent after it; Figure 9 illustrates this directly, showing the same template used to reference Area 51 in July 2019 and Covid-19 in December 2020. Notably, despite the template imposing no formal constraint on text, creators spontaneously converge on the construction *XX Everywhere*, with only the variable slot shifting across topics. Further examples of such conventionalized expressions are provided in Appendix A.7.

Template 3,109 shows a similar pattern of drift around culturally salient references, though with a notable distinction. The term *september 20th* references the Storm Area 51 event, a real news that captured widespread attention in 2019 (Figure 9f). In contrast, a surge of posts in January 2022 reference former Vice President Dick Cheney suffering a heart attack on June 17th, 2022 (Figure 9h). This date was still five months in the future at the time of posting, and the event never occurred, highlighting how memes can serve as vehicles for the spread of misinformation.

6 Conclusion and Future Work

We present MEMESOVERTIME, a large-scale dataset of 6.6M meme instances collected from Reddit over roughly a decade, grounded in the messy, diverse, and ever-changing landscape of real-world online meme culture. We curate MEMESOVERTIME through a semi-automatic pipeline combining two-stage clustering with manual consolidation, defining templates by both visual consistency and preservation of meaning across instances.

Using MEMESOVERTIME, we conduct a temporal analysis of meme template activity along multiple dimensions. We find that template activity is front-loaded: most templates reach peak usage within the first third of their active span, with shorter-lived templates concentrating nearly all their activity in the opening months. Among sustained-large templates, those with more instances show less lexical drift, plausibly because their broader vocabulary makes newly introduced terms shift the overall profile less.

We hope MEMESOVERTIME serves as a foundation for future research on the temporal dynamics of internet culture through memes, including the spread of meme templates across communities, or the role of visual templates in shaping political discourse.

7 Limitations

Tracking the long-term evolution of memes is the central motivation of this work, and achieving this at scale ideally requires automated template identification. We explored this direction, but encountered failures that we believe reflect fundamental properties of memes themselves rather than merely methodological limitations.

The Open-Ended Nature of Templates Meme templates emerge continuously and unpredictably. Any database-driven approach to template identification is therefore inherently incomplete: a template that does not yet exist at annotation time will remain unrecognized, causing novel instances to be misclassified or discarded. This is not a limitation that more data alone can resolve, as the space of possible templates is effectively unbounded.

Visual Entity vs. Template Structure Embedding-based methods such as CLIP clustering tend to latch onto the most visually salient features of an image, which are not always the features that define the template. In the *Drake*

Pointing template, for instance, the defining structure is the two-panel approval/disapproval format. When the figure of Drake is replaced by a celebrity or fictional character, CLIP redirects its attention to the substituted figure, grouping the meme with visually similar images rather than with other instances of the same template. The instantiation variable, in other words, overwhelms the template frame in the embedding space.

Template Remixing A subtler challenge arises when two recognizable templates are combined. Here the difficulty is not that a visual entity competes with a template, but that two templates compete with each other for primacy. It becomes unclear which element constitutes the stable, reusable frame and which has been absorbed as a variable. Such cases resist clean categorization and expose the limits of any scheme that treats template identity as a fixed, discrete property.

In light of these challenges, we adopt a semi-automatic approach in this work. We hope that the manual annotations provided in MemeAtlas can serve as a foundation for future efforts to address these difficulties, including automated detection of emerging templates, large-scale template classification, and a deeper understanding of how meme templates generalize across visual instantiations.

Ethics Statement

All annotations were conducted in accordance with ethical guidelines, ensuring that annotators were not exposed to any psychologically distressing content during the process. All annotators are paid according to national standards.

Use of AI Assistants. The authors acknowledge the use of Claude for correcting grammatical errors, enhancing the coherence of the final manuscripts, and providing assistance with coding.

Acknowledgements

We thank MaiNLP lab members for giving feedback on earlier drafts of this paper, as well as to the anonymous reviewers for their feedback. This work is supported by the KLIMA-MEMES project funded by the Bavarian Research Institute for Digital Transformation (bidt), an institute of the Bavarian Academy of Sciences and Humanities. The authors are responsible for the content of this publication.

References

- Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhi-fang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. [Qwen3-vl technical report](#). *Preprint*, arXiv:2511.21631.
- Kate Barnes, Péter Juhász, Marcell Nagy, and Roland Molontay. 2024. [Topicality boosts popularity: A comparative analysis of NYT articles and Reddit memes](#). *Social Network Analysis and Mining*, 14(1):119.
- Kate Barnes, Tiernon Riesenmy, Minh Duc Trinh, Eli Lleshi, Nóra Balogh, and Roland Molontay. 2021. [Dank or not? Analyzing and predicting the popularity of memes on Reddit](#). *Applied Network Science*, 6(1):21.
- Christian Bauckhage. 2021. [Insights into internet memes](#). *Proceedings of the International AAAI Conference on Web and Social Media*, 5(1):42–49.
- Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. 2020. [The pushshift reddit dataset](#). *ArXiv*, abs/2001.08435.
- Lars Bülow, Marie-Luis Merten, and Michael Johann. 2018. [Internet-Memes als Zugang zu multimodalen Konstruktionen](#). *Zeitschrift für Angewandte Linguistik*, 2018(69):1–32.
- Viktor Chagas, Fernanda Freire, Daniel Rios, and Dan-dara Magalhães. 2019. [Political memes and the politics of memes: A methodological proposal for content analysis of online political memes](#). *First Monday*.
- Cheng Cui, Ting Sun, Manhui Lin, Tingquan Gao, Yubo Zhang, Jiaxuan Liu, Xueqing Wang, Zelun Zhang, Changda Zhou, Hongen Liu, Yue Zhang, Wenyu Lv, Kui Huang, Yichao Zhang, Jing Zhang, Jun Zhang, Yi Liu, Dianhai Yu, and Yanjun Ma. 2025. [Paddleocr 3.0 technical report](#). *Preprint*, arXiv:2507.05595.
- Barbara Dancygier and Lieven Vandelanotte. 2017. [Internet memes as multimodal constructions](#). *Cognitive Linguistics*, 28(3):565–598.
- Richard Dawkins. 1976. *The Selfish Gene*. Oxford University Press, New York and Oxford.
- Shiling Deng, Serge Belongie, and Peter Ebert Christensen. 2025. [Large Vision-Language Models for Knowledge-Grounded Data Annotation of Memes](#).
- EunJeong Hwang and Vered Schwartz. 2023. [MemeCap: A Dataset for Captioning and Interpreting Memes](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1433–1445, Singapore. Association for Computational Linguistics.
- Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020. [The Hateful Memes Challenge: Detecting Hate Speech in Multimodal Memes](#). In *Advances in Neural Information Processing Systems*, volume 33, pages 2611–2624. Curran Associates, Inc.
- Benjamin Lambright, Jordan Youner, and Constantine Lignos. 2025. [MetaMeme: A Dataset for Meme Template and Meta-Category Classification](#). In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop)*, pages 356–367, Albuquerque, USA. Association for Computational Linguistics.
- Adam Lonnberg, Pengcheng Xiao, and Kathryn Wolfinger. 2020. [The growth, spread, and mutation of internet phenomena: A study of memes](#). *Results in Applied Mathematics*, 6:100092.
- Leland McInnes, John Healy, and Steve Astels. 2017. [hdbscan: Hierarchical density based clustering](#). *Journal of Open Source Software*, 2(11):205.
- Leland McInnes, John Healy, and James Melville. 2020. [Umap: Uniform manifold approximation and projection for dimension reduction](#). *Preprint*, arXiv:1802.03426.
- Levente Murgás, Marcell Nagy, Kate Barnes, and Roland Molontay. 2024. [Decoding memes: A comparative study of machine learning models for template identification](#). *Preprint*, arXiv:2408.08126.
- Andrew Ng, Michael Jordan, and Yair Weiss. 2001. [On spectral clustering: Analysis and an algorithm](#). In *Advances in Neural Information Processing Systems*, volume 14. MIT Press.
- Khoi P. N. Nguyen and Vincent Ng. 2024. [Computational Meme Understanding: A Survey](#). In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21251–21267, Miami, Florida, USA. Association for Computational Linguistics.
- Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, and 7 others. 2024. [Dinov2: Learning robust visual features without supervision](#). *Preprint*, arXiv:2304.07193.
- Shraman Pramanick, Shivam Sharma, Dimitar Dimitrov, Md. Shad Akhtar, Preslav Nakov, and Tan-moy Chakraborty. 2021. [MOMENTA: A Multimodal Framework for Detecting Harmful Memes and Their Targets](#). In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4439–4455,

- Punta Cana, Dominican Republic. Association for Computational Linguistics.
- 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. 2021. [Learning Transferable Visual Models From Natural Language Supervision](#). In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.
- Ursula Kristin Schmid, Heidi Schulze, and Antonia Drexel. 2025. [Memes, humor, and the far right’s strategic mainstreaming](#). *Information, Communication & Society*, 28(4):537–556.
- Chhavi Sharma, Deepesh Bhageria, William Scott, Srinivas PYKL, Amitava Das, Tanmoy Chakraborty, Viswanath Pulabaigari, and Björn Gambäck. 2020. [SemEval-2020 Task 8: Memotion Analysis- the Visuo-Lingual Metaphor!](#) In *Proceedings of the Fourteenth Workshop on Semantic Evaluation*, pages 759–773, Barcelona (online). International Committee for Computational Linguistics.
- Limor Shifman. 2014. *Memes in Digital Culture*. MIT Press.
- Riccardo Tommasini, Filip Ilievski, and Thilini Wijesiriwardene. 2023. [IMKG: The Internet Meme Knowledge Graph](#). In *The Semantic Web: 20th International Conference, ESWC 2023, Hersonissos, Crete, Greece, May 28–June 1, 2023, Proceedings*, pages 354–371, Berlin, Heidelberg. Springer-Verlag.
- Carlo M. Valensise, Alessandra Serra, Alessandro Galeazzi, Gabriele Etta, Matteo Cinelli, and Walter Quattrociocchi. 2021. [Entropy and complexity unveil the landscape of memes evolution](#). *Scientific Reports*, 11(1):20022.
- Camilla Vásquez and Erhan Aslan. 2021. [“Cats be outside, how about meow”: Multimodal humor and creativity in an internet meme](#). *Journal of Pragmatics*, 171:101–117.
- Savvas Zannettou, Tristan Caulfield, Jeremy Blackburn, Emiliano De Cristofaro, Michael Sirivianos, Gianluca Stringhini, and Guillermo Suarez-Tangil. 2018. [On the Origins of Memes by Means of Fringe Web Communities](#). In *Proceedings of the Internet Measurement Conference 2018, IMC ’18*, pages 188–202, New York, NY, USA. Association for Computing Machinery.
- Eline Zenner and Dirk Geeraerts. 2018. [One does not simply process memes: Image macros as multimodal constructions](#). In Esme Winter-Froemel and Verena Thaler, editors, *Cultures and Traditions of Wordplay and Wordplay Research*, pages 167–194. De Gruyter.
- Naitian Zhou, David Jurgens, and David Bamman. 2024. [Social Meme-ing: Measuring Linguistic Variation in Memes](#). In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3005–3024, Mexico City, Mexico. Association for Computational Linguistics.
- Shijia Zhou, Saif M. Mohammad, Barbara Plank, and Diego Frassinelli. 2026. [I came, I saw, I explained: Benchmarking multimodal LLMs on figurative meaning in memes](#). In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 9460–9477, Palma de Mallorca, Spain. ELRA Language Resource Association.
- Shijia Zhou, Siyao Peng, Simon M. Luebke, Jörg Haßler, Mario Haim, Saif M. Mohammad, and Barbara Plank. 2025. [What media frames reveal about stance: A dataset and study about memes in climate change discourse](#). In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 5337–5356, Suzhou, China. Association for Computational Linguistics.

Appendix

A MEMESOVERTIME Curation Details

A.1 Image Preprocessing Details

Text masking We detected text regions with PaddleOCR (Cui et al., 2025), and masked them with the median color of surrounding pixels.

Margin removal Margins are removed by cropping the outermost border pixels until no uniform-color margin remains.

A.2 Clustering Implementation Details

We encode each preprocessed image using CLIP ViT-L/14 (Radford et al., 2021) to obtain embeddings $\mathbf{v}_i \in \mathbb{R}^{768}$, which are ℓ_2 -normalised:

$$\hat{\mathbf{v}}_i = \frac{\mathbf{v}_i}{\|\mathbf{v}_i\|_2} \quad (2)$$

To reduce noise and computational cost, we draw a month-stratified 5% sample ($\approx 950\text{K}$ images) and reduce their embeddings $\hat{\mathbf{v}}_i$ to 20 dimensions via UMAP (McInnes et al., 2020):

$$\mathbf{u}_i = \text{UMAP}(\hat{\mathbf{v}}_i) \in \mathbb{R}^{20} \quad (3)$$

We then apply HDBSCAN (McInnes et al., 2017) to the reduced embeddings $\mathbf{u}_i$ to obtain cluster labels ℓ_i . Since HDBSCAN only labels the sample, we propagate labels to the remaining instances via k -nearest neighbour ($k=5$) voting in the reduced space:

$$\ell_i = \text{mode}_c \{ \ell_j : j \in \mathcal{N}_k(i) \} \quad (4)$$

where $\mathcal{N}_k(i)$ denotes the k nearest neighbours of point i among the HDBSCAN-labelled sample points.

A.3 Subclustering Implementation Details

Since images within each primary cluster already share a degree of visual similarity, their DINOv2 embeddings occupy a relatively compact region of the embedding space. We therefore apply PCA to reduce dimensionality prior to Spectral Clustering (Ng et al., 2001), retaining 70% of variance, which removes noise while preserving the local structure relevant for fine-grained visual discrimination. Spectral Clustering is then applied as it makes no assumptions about cluster geometry and is well-suited to the non-convex structure typical of visual embedding spaces. The number of neighbours for the affinity graph is set adaptively as a function of cluster size and mean pairwise cosine distance.

A.4 Annotation Guidelines and Examples

Annotation proceeded in two stages. In the first stage, 2 annotators performed intra-cluster consolidation, filtering out non-meme content (Figure 10) and merging instances that share the same underlying template (Figure 11). In the second stage, 2 annotators performed inter-cluster consolidation, comparing pairs of clusters using the same merging criteria.

Task: Cluster Denoising

You will be shown 10 images from a cluster. Please answer the following questions:

Q1. Are these images memes?

The following are examples of content that is *not* considered a meme (this list is not exhaustive): QR codes, screenshots of search engine results, game interface captures, online chat screenshots, and images consisting solely of lyrics or quotes overlaid on a background.

Q2. Can you identify at least 8 images that share the same meme template?

To determine whether two images share the same meme template, apply the following criteria in order:

1. **Base image consistency.** Imagine removing all user-added text and edits. Is the underlying base image the same?

- *Yes* → Same template.
- *Not identical, but structurally similar* with a clear correspondence between elements → Proceed to criterion 2.
- *Unrelated* → Not the same template.

2. **Meaning consistency.** Imagine swapping the user-added content between the two images into their corresponding positions. Do both memes still convey the same meaning?

- *Yes* → Same template.
- *No* → Not the same template.

Figure 10: Annotation guidelines for cluster denoising.

Task: Cluster Merging

You will be shown two rows of images, each containing 10 examples from a different cluster. Compare the dominant meme template in each row and determine whether they share the same underlying template, using the following criteria:

1. **Base image consistency.** Imagine removing all user-added text and edits. Is the underlying base image the same?
 - *Yes* → Same template.
 - *Not identical, but structurally similar* with a clear correspondence between elements → Proceed to criterion 2.
 - *Unrelated* → Not the same template.
2. **Meaning consistency.** Imagine swapping the user-added content between the two images into their corresponding positions. Do both memes still convey the same meaning?
 - *Yes* → Same template.
 - *No* → Not the same template.

Figure 11: Annotation guidelines for cluster merging.

Figure 12 shows instances of the same template at varying degrees of cropping; Figure 13 shows visually similar images featuring the similar feature (big eyes) but drawn from distinct templates; Figure 14 shows a mixed cluster containing multiple templates that primary clustering failed to separate; and Figure 15 shows noisy instances automatically grouped together, isolating them from coherent template clusters.

A.5 Inter-annotator Agreement

Both annotation steps were carried out by two in-house annotators: the first author and a Master’s student in Computational Linguistics. Both annotators were hired and compensated according to national standards. During inter-cluster consolidation, annotators jointly labeled 1,200 instances, achieving a Cohen’s κ of 0.87.

A.6 OCR Extraction and Watermark Removal

Each image in MEMESOVERTIME is processed by two OCR pipelines: PaddleOCR (Cui et al., 2025) and Qwen3-VL⁷ (Bai et al., 2025).

⁷<https://huggingface.co/Qwen/Qwen3-VL-30B-A3B-Instruct>

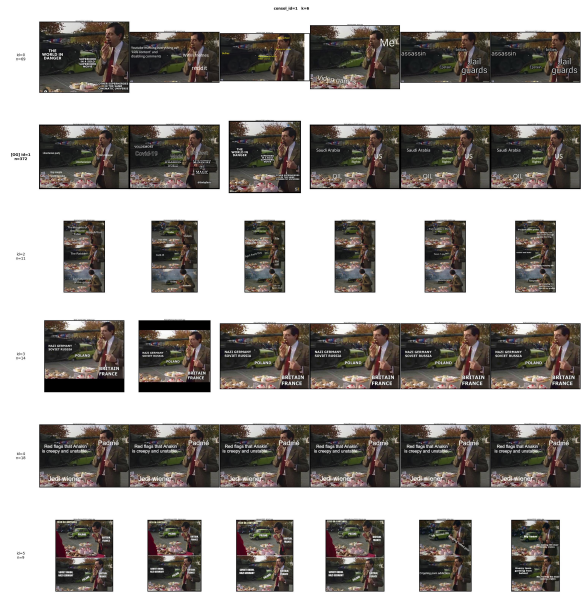


Figure 12: Cropping variation.



Figure 13: Similar feature, different templates.

For Qwen3-VL, we extract the OCR with the following prompt:

Please OCR this image. Read all text from left to right, top to bottom, and output every character exactly as it appears.

Qwen3-VL was deployed on a single NVIDIA H200 140GB GPU, with inference completing in 186 hours.

To evaluate their complementary strengths, we manually compared outputs on a sample of 100 memes and found that Qwen3-VL consistently produces better-structured output, in particular:

- **Line segmentation:**

post_id: b83rqg

PaddleOCR: “Nobody: ElonMusk: Hear Elon Musk’s SurpriseRap Song ‘RIP Harambe’ rollingstone.com”

Qwen3-VL: “Nobody: / Elon Musk: / Hear Elon Musk’s Surprise Rap Song ‘RIP Harambe’ / rollingstone.com”

- **Emoji and token preservation:**

post_id: 1i5lpqr

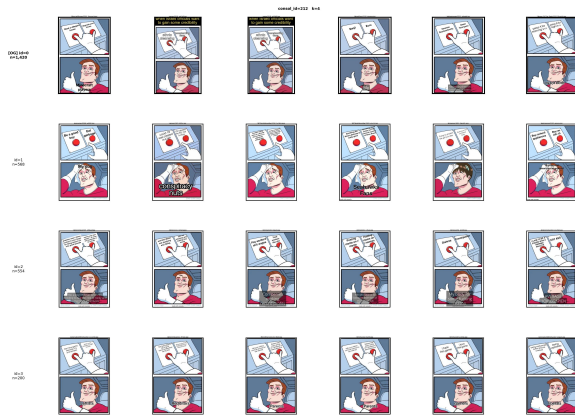


Figure 14: Mixed templates.

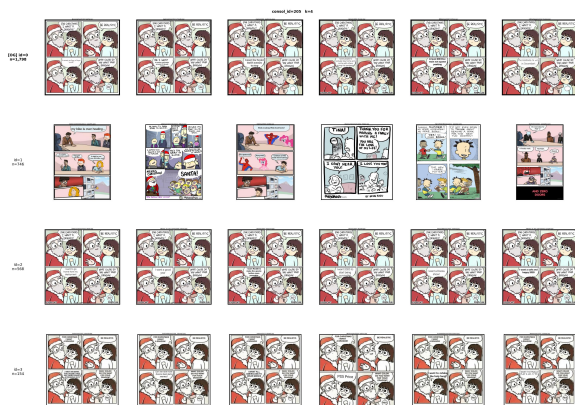


Figure 15: Noise cluster.

PaddleOCR: “Who made this? TheCovid-19VaccineOnMySon Testing”

Qwen3-VL: “Who made this? [crying emoji] / Testing The Covid-19 Vaccine On My Son”

We therefore retain both fields—`paddle_ocr` for its higher coverage and `qwen_ocr` for its structural quality—as complementary text representations in MEMESOVERTIME.

Since memes are frequently created using third-party platforms (e.g., Imgflip, Mematic, Meme-droid), OCR output often contains platform watermarks appended to the image text. We apply a rule-based procedure targeting known signatures (e.g., `imgflip.com`, `made with mematic`, `JAKE-CLARK.TUMBLR`) at the *end* of the OCR string, applied iteratively until convergence.

Watermarks are deliberately restricted to end-of-string matches, as some memes organically discuss these platforms (e.g., complaining about watermarks; see Figures 16a–16c). Residual occurrences (<0.01%) stem from OCR corruption or such organic references, where removal would corrupt legitimate content.



(a) post_id: cykgbc
“Using mematic / Not knowing how to make a meme without a watermark”



(b) post_id: c2r9n6
“Me putting memes on Imgflip / Then realizing I can’t remove the template”



(c) post_id: x4zbnk
“society if the mematic watermark was gone:”



(d) post_id: b93kx5
“Using editing software / Using ImgFlip / Using mematic / Using Google Slide”

Figure 16: Examples of meme posts whose OCR text organically references meme creation platforms. Removing platform-related tokens from these posts would corrupt legitimate semantic content.

A.7 Conventionalized Expressions

Conventionalized expressions are not an isolated phenomenon in MEMESOVERTIME. Figure 17 shows ten instances of Template 569, whose creators consistently adopt the construction *I go to X*. Figure 18 shows ten instances of Template 236, associated with the expression *I am speed*, derived from the film *Cars*.

These expressions can become sufficiently entrenched that the template as a whole circulates as a cultural reference in its own right. Figure 19 illustrates this: the *I am speed* template, image and text together, is embedded wholesale into a different template, treating one meme template as the content of another.

B Robustness to Observation-Window Censoring

Our corpus covers a fixed window from January 2016 to December 2025, which raises the question of whether the temporal patterns reported in Section 5.1 are artifacts of that window rather than properties of the templates themselves. To address this, we group templates by the year in which they first appear and measure, for each year, the number of templates, the median observable window (months from first appearance to the 2025-12 cut-

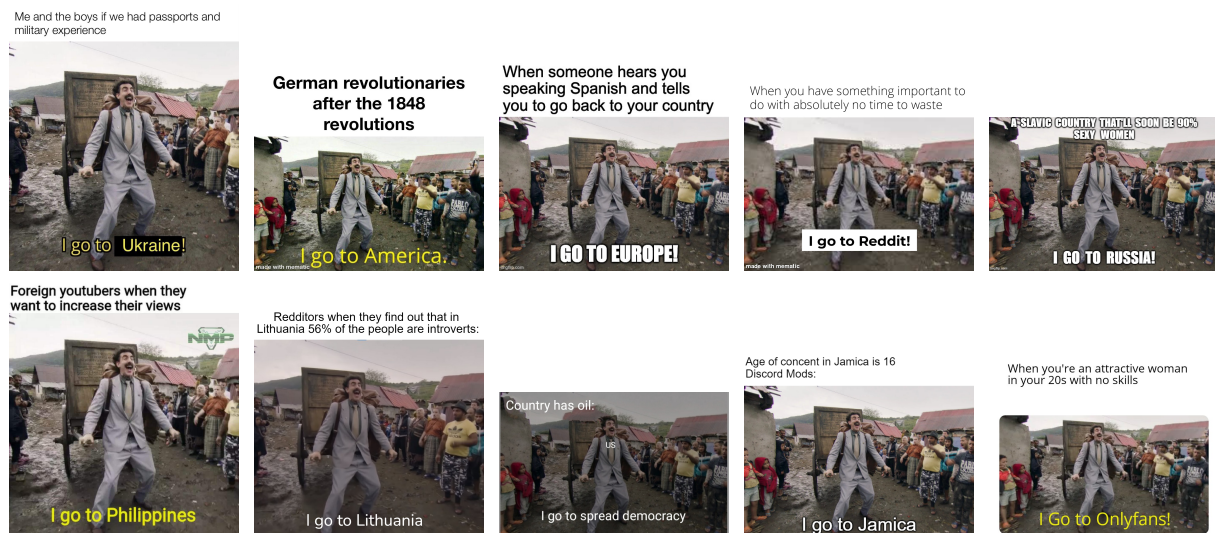


Figure 17: Ten instances of Template 569, illustrating the conventionalized construction *I go to X*.

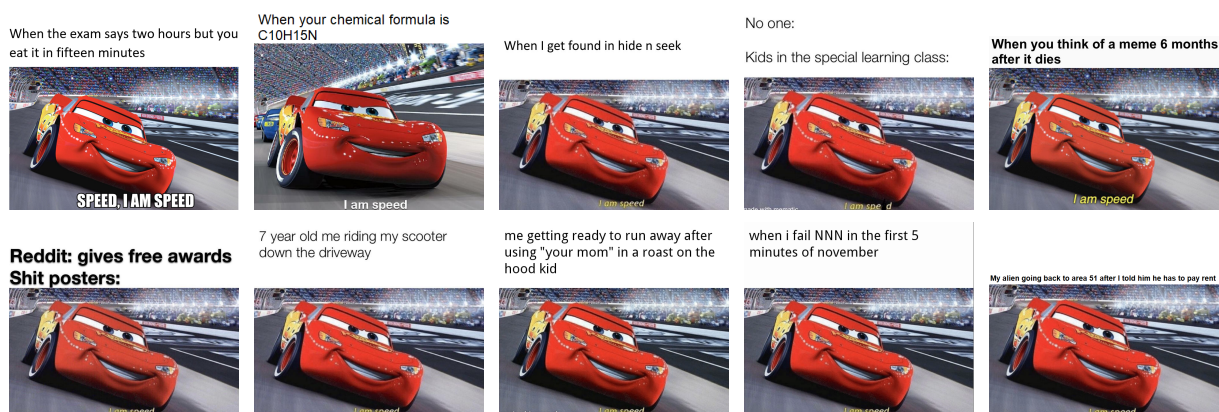


Figure 18: Ten instances of Template 236, illustrating the conventionalized expression *I am speed*.



Figure 19: Examples of *I am speed* used as a cultural reference within a different template, independent of its origin in Template 236.

off), the median number of active months, and the median number of instances per template.

Our measures are not driven by template age. Among templates that first appeared between 2019 and 2023, the median observable window shrinks

from 79 to 29 months, yet the median number of active months (27, 26, 26, 28, 25) and the median number of instances (117, 102, 113, 101, 101) stay essentially flat. If either measure depended on how long a template has been around, it would follow the observable window and drop along with it. Neither does, which indicates that active months and template size capture properties of template usage rather than elapsed observation time. The higher values for the 2016 cohort fit a survivorship pattern rather than an age effect: the older templates still identifiable today tend to be the genuinely large, long-lived ones.

Right-censoring. Templates that first appear late in the observation window are cut off at the end, so their peak may fall outside our data. Exposure to this form of censoring is small: 91.4% of templates first appear before 2020, 96.8% have an observable window of at least 60 months, and only 4 templates (0.1%) have a window shorter than 12 months. Re-

running the peak-timing analysis of Section 4 on mature templates only, that is, those with an observable window over 42 months, leaves the early-peak finding unchanged.

Left-censoring. Our data begins in January 2016 because earlier Reddit meme data is too sparse for reliable analysis (Section 4). A template that originated before 2016 therefore has its true onset, and possibly its true peak, outside our window, and what we observe is the later portion of its lifespan. This works against the pattern we report: losing the early part of a trajectory can only shift an estimated peak toward the end of the observed period, never toward the beginning. Left-censoring can dampen or obscure an early-peak pattern, but it cannot manufacture one. The early-peak finding is therefore conservative with respect to this source of bias.