BigTokDetect: A Clinically-Informed Vision–Language Model Framework for Detecting Pro-Bigorexia Videos on TikTok

Anonymous ACL submission

Abstract

Social media platforms increasingly struggle to detect harmful content that promotes muscle dysmorphic behaviors, particularly pro-bigorexia content that disproportionately affects adolescent males. Unlike traditional eating disorder detection focused on the "thin ideal," pro-bigorexia material masquerades as legitimate fitness content through complex multimodal combinations of visual displays, coded language, and motivational messaging that evade text-based detection systems. We address this challenge by developing BIGTOKDETECT, a clinically-informed detection framework for identifying pro-bigorexia content on TikTok. We introduce BigTok, the first expert-annotated multimodal dataset of over 2,200 TikTok videos labeled by clinical psychologists and psychiatrists across five primary categories spanning body image, nutrition, exercise, supplements, and masculinity. Through comprehensive evaluation of state-of-the-art vision language models, we achieve 0.829% accuracy on primary category classification and 0.690% on subcategory detection via domain-specific finetuning. Our ablation studies demonstrate that multimodal fusion improves performance by 5-10% over text-only approaches, with video features providing the most discriminative signals. These findings establish new benchmarks for multimodal harmful content detection and provide both the computational tools and methodological framework needed for scalable content moderation in specialized mental health domains.

1 Introduction

Social media platforms face mounting challenges in detecting harmful content that significantly impacts user mental health (Chancellor and Choudhury, 2020; Gorwa et al., 2020). Contemporary platforms like TikTok algorithmically amplify content promoting unrealistic body ideals (Becker, 2004; Minadeo and Pope, 2022), yet much harmful material exists in gray areas where legitimate discussions intersect with dangerous messaging (Gillespie, 2018). These narratives emerge through complex multimodal combinations—visual imagery, audio cues, and textual descriptions—that traditional text-based detection systems cannot adequately capture (Kiela et al., 2020a,b). Effective automated moderation requires sophisticated multimodal approaches that can parse nuanced signals across video, audio, and text (Gimeno-Gómez et al., 2024).

Pro-bigorexia content—material promoting muscle dysmorphic behaviors—exemplifies these multimodal detection challenges at their most complex. This harmful content promotes compulsive muscle-building behaviors that disproportionately affect adolescent males (Pope Jr et al., 1997; Mitchison et al., 2022), yet remains severely understudied in computational research compared to traditional eating disorder detection (Chancellor and Choudhury, 2020). Pro-bigorexia material masquerades as legitimate fitness content through subtle combinations of muscular displays, extreme workout demonstrations, and coded supplement language (Murray et al., 2017; Kamkari, 2025). The visual similarity between harmful and beneficial fitness content, combined with evolving coded terminology, renders current text-based detection approaches fundamentally inadequate.

Current automated detection systems face critical limitations when confronting such content. Foundation models trained on general corpora lack domain-specific knowledge to recognize subtle clinical markers and euphemistic language patterns (Murray et al., 2017). Existing systems rely heavily on keyword filtering and text-based signals, missing critical visual and behavioral cues embedded in video content (Kiela et al., 2020b; Gorwa et al., 2020). The dynamic nature of social media—where creators deliberately evolve language to evade detection—further challenges

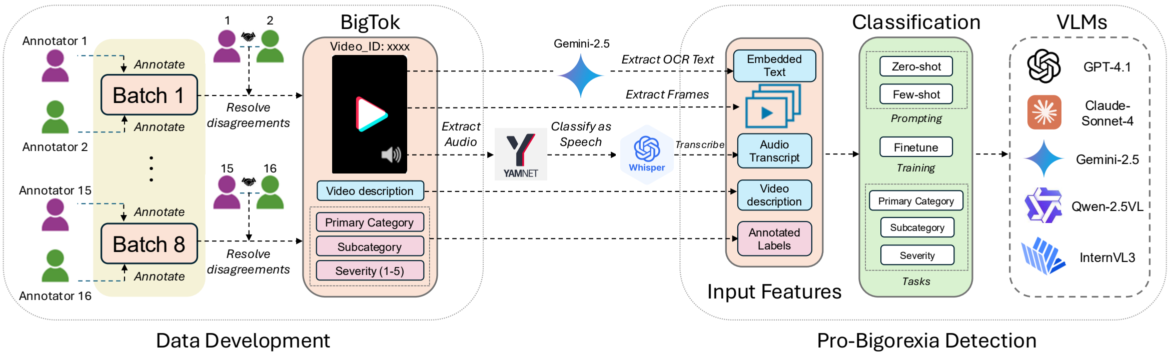

Figure 1: BIGTOK pipeline overview. Left: Expert annotation process with dual annotation and consensus resolution. Right: Multimodal feature extraction (visual, audio, text) and classification evaluation across VLMs using zero-shot, few-shot, and finetuning approaches for primary category, subcategory, and severity prediction tasks.

automated approaches (Gillespie, 2018). Effective detection requires integrated analysis of visual behavioral cues, audio transcripts, and textual context—capabilities that remain underexplored in computational mental health research (Gimeno-Gómez et al., 2024).

To address these detection challenges, we make the following contributions:

- We develop the first clinically-informed taxonomy for automated pro-bigorexia detection, establishing fine-grained categories spanning body image, nutrition, supplement abuse, exercise practices, and masculinity that enable systematic computational analysis.
- We introduce BIGTOK, a multimodal dataset of over 2,200 expert-annotated TikTok videos that enables robust evaluation of vision-language models on challenging harmful content detection tasks.
- We construct BIGTOKDETECT, a detection framework that achieves state-of-the-art (SOTA) performance using proprietary and open-source VLMs, with our best models reaching 95.1% accuracy on primary category classification and 91.0% on fine-grained subcategory detection through fine-tuning.
- We demonstrate through comprehensive ablation studies that multimodal fusion improves detection performance by 5-10% over text-only approaches, with video features providing the most discriminative signals for identifying subtle pro-bigorexia behaviors.

Figure 1 summarizes our comprehensive approach to multimodal pro-bigorexia detection. Our methodology demonstrates that clinically-informed vision-language models can effectively identify subtle harmful content that evades traditional detection systems. Through systematic evaluation across multiple model architectures and training paradigms, we achieve substantial improvements over existing text-based methods, establishing a new paradigm for detecting nuanced harmful content on social media platforms. The BIGTOKDETECT framework provides both the computational tools and replicable methodology needed to address the growing challenge of automated content moderation in mental health domains.

2 Related Work

Muscle Dysmorphic Disorder on Social Media Muscle dysmorphic disorder, or "bigorexia," involves preoccupation with insufficient muscularity, driving compulsive behaviors including exercise, rigid dieting, and supplement abuse (Pope Jr et al., 1997; Cooper et al., 2020). Frequent exposure to muscularity-oriented TikTok and Instagram content—including body-transformation videos, supplement promotions, and steroid use—has been linked to elevated rates of probable muscle dysmorphia (Mitchison et al., 2022; Ganson et al., 2025). Hypermasculine online subcultures (e.g., the "manosphere") intensify these pressures through toxic social comparisons and body-optimization trends (Kamkari, 2025). Despite this prevalence, both psychological research and platform moderation remain focused on anorexia and bulimia, leaving bigorexia largely unmoderated (Lookingbill et al., 2023).

Multimodal Mental Health and Body Image Content Detection Automated detection of body image content has overwhelmingly targeted anorexia and bulimia (Chancellor et al., 2017; Chancellor and Choudhury, 2020), relying primarily on keyword filters and text-only classifiers despite evidence that vision-language fusion substantially improves detection of subtle harmful imagery (Kiela et al., 2020b). Short-form video platforms have become dominant for youth mental health discourse (Basch et al., 2022), yet eating disorder detection remains predominantly text-based (Wang et al., 2017; Merhbene et al., 2024). Even dedicated TikTok corpora for eating disorders research exclude muscle dysmorphia labels (Bickham et al., 2025; Donati et al., 2023), while existing qualitative analyses of pro-muscularity forums (Murray et al., 2015) lack the scale needed for robust computational detection. Modern vision-language models—Flamingo (Alayrac et al., 2022), InstructBLIP (Li et al., 2023), GPT-4V (OpenAI, 2023)—enable end-to-end multimodal detection but lack clinically grounded training data for specialized tasks. Existing mental health datasets rely on crowdsourced annotations rather than clinical expertise, potentially missing subtle markers needed to distinguish legitimate fitness content from harmful pro-bigorexia messaging that uses coded language around supplements and steroids (Murray et al., 2017). To our knowledge, no large-scale expert-annotated datasets exist for this increasingly prevalent content type.

3 BigTok: Expert-Curated Multimodal Dataset Construction

3.1 Data Collection

We source our corpus from TikTok via its official API (TikTok, 2025). We query videos from January 2019 to January 2025 to capture pre- and post-pandemic trends. Guided by domain experts, we curate 40 high-precision query terms mapped to taxonomy subcategories (Table 12, Appendix A.1), retrieving up to 1,000 videos and their metadata per term. We randomly sample videos per group of keywords that belong to each primary and subcategory to 2,400 videos for annotation. Only video content and captions are exposed to annotators to protect user privacy (anonymization steps are described in Appendix C). We also compile negative (irrelevant) examples using 42 trending hashtags from TikTok's Creative Center (TikTok Creative Center), following established supervised classification practices (Kiela et al., 2020b).

3.2 Annotating Bigorexia Content

Annotator Recruitment and Demographics We recruit 16 subject matter experts (13 females and 3 males), including licensed clinical psychologists, social workers, and doctoral candidates with research and clinical experience in eating and body image disorders (Table 11, Appendix B.1 for profiles). Their combined expertise in empirical research and direct patient care ensures that our annotations are both theoretically grounded and have clinical relevance.

Annotation Instructions We base our annotation interface on Amazon Mechanical Turk via an invitation-only pool (Appendix B.3). Annotators are instructed to select the first (mandatory) primary–subcategory pair ($[t_1, s_1]$) and the second (optional) primary–subcategory pair ($[t_2, s_2]$). Every video is also rated for harm on a discrete 5-point Likert scale with 0.5-point increments (1=not harmful and 5=very harmful). The annotation interface is shown in Appendix B.2.

Annotator Training As a pilot study, each annotator independently labels a set of 20-30 videos to validate taxonomy coverage, identify potential edge cases, and familiarize themselves with the annotation interface. We monitor disagreements and encourage detailed note-taking. Annotators then join a group session to discuss flagged issues, collaboratively annotate selected videos while explaining their reasoning, until consensus is reached.

Batchwise Annotation We split the dataset into 8 batches of ~300 videos. For each batch, we filter out videos that are not available (e.g., they have been deleted) or are not in English. Then the batch is assigned to two different annotators, so that every video receives two independent annotations. To assess agreement between a pair of annotators A and B, we compare all combinations of labels and determine the level of agreement:

- **Perfect agreement:** both the first and second primary–subcategory pairs match exactly across annotators, i.e. $\forall i \in \{1, 2\} : [t_i, s_i]^A = [t_i, s_i]^B$
- **Strong agreement:** one primary–subcategory pair matches exactly across annotators, i.e. $\exists i \in \{1, 2\} : [t_i, s_i]^A = [t_i, s_i]^B$
- **Weak agreement:** at least one primary category matches across any labels, but the subcategories differ, i.e. $\exists i \in \{1, 2\} : t_i^A = t_i^B \wedge s_i^A \neq s_i^B$

- **Disagreement:** no common primary category across labels, i.e. $\forall i \in \{1, 2\},\ t_i^A \neq t_i^B \ \wedge\ s_i^A \neq s_i^B$.

We then flag examples with weak agreement or disagreement on pro-bigorexia detection, and examples where severity scores differ by more than two points. The two annotators are invited to a structured discussion to articulate their reasoning, review cases that need consultation, address disagreements, and eventually aim to reach consensus, although consensus is not mandatory. After the discussion, each annotator independently re-annotates the flagged videos. Further details on this procedure are provided in Appendix B.4.

Label Aggregation While each video may receive multiple label pairs (primary-subcategory), for sake of easy evaluation, we assign a single label pair to each video. For videos with perfect agreement, we randomly choose of the label pairs. For strong agreement, we adopt the shared primary–subcategory pair; for weak agreement, we choose the mandatory (first) label over the optional (second) one; remaining disagreements are resolved by random tie-breaking. The final harm severity score is the average of the two annotators' ratings. This simplification reflects current limitations of LLMs in reliable multi-label classification (Ma et al., 2025).

We assess inter-rater reliability with Cohen's κ (Cohen, 1960) for two annotators. Initial agreement is moderate (strong $\kappa = 0.43$–0.69; weak $\kappa = 0.59$–0.83; Table 6). After structured consensus discussions and re-annotation, reliability increases substantially (strong $\kappa = 0.58$–0.81; weak $\kappa = 0.78$–0.94), ensuring that our iterative process yields high-quality, consistent labels. All Cohen's κ values for each batch are shown in Table 6 (Appendix B.5)

3.3 Data Statistics

Our final BIGTOK dataset comprises 2,210 annotated TikTok videos. The median video duration is 20.6 seconds, and 59.8% of videos are under 30 seconds. Table 1 shows the label distribution. At the primary category level, excluding *Irrelevant*, *Relationship to Exercise* (450) is the most dominant class. At the subcategory level, *Muscularity Self-objectification* predominates (278), reflecting the popularity of the genre of videos in which people display their muscles. The *Other* subcategory covers videos within a primary category that do not match predefined subcategories (e.g., humor or lifestyle clips); these are excluded from subcategory-level benchmarking. Some examples from BIGTOK are shown in Tables 13, 14, and 15 Appendix A.1. Additional annotation statistics are provided in Appendix B.5.

Primary Category	**Count**
Relationship to Body	449
Relationship to Exercise	450
Relationship to Food	317
Supplement Abuse	220
Relationship to Masculinity	131
Irrelevant	639

Subcategory	**Count**
Muscularity Self-objectification	278
Leanness Self-objectification	72
Muscle Dissatisfaction	78
Rigid Food Rules	136
Unsafe Food	79
Cheat Meals	74
Excessive Exercise	109
Predebting Exercise	22
Maladaptive Coping	89
Exercise-Induced Functional Impairment	34
Toxic Motivation	107
Anabolic Steroids	68
Legal APEDs	50
Hormone Therapy	82
Relationship to Masculinity	131
Other	103
Irrelevant	639

Table 1: Distribution of categories and subcategories BIGTOK. We include the primary categories *Irrelevant* and *Relationship to Masculinity*, which can also be considered as subcategories.

4 Classifying Multimodal Pro-Bigorexia Content via VLMs

To develop an automated detection system for pro-bigorexia content, we apply SOTA VLMs to this challenging multimodal classification task. We leverage both commercial and open-source VLMs through zero-shot prompting, few-shot learning, and fine-tuning approaches to achieve high performance in identifying subtle pro-bigorexia behaviors across video, audio, and text modalities.

4.1 Task Definition

We define three evaluation tasks for TikTok video classification. **Task 1: Primary Category Classification** involves predicting one of the five primary categories of pro-bigorexia content (*Body*, *Food*, *Supplements*, *Exercise*, *Masculinity*), or *Irrelevant*. **Task 2: Subcategory Classification** involves predicting the specific subcategory within the selected primary category. **Task 3: Severity Estimation** involves predicting a continuous severity score on a 1-5 scale, where 1 indicates no harm and 5 indicates extreme harm.

For each task, we split the data into training and test sets at a 3:1 ratio using stratified sampling. The test set is strictly balanced across all classes, while the training set is adjusted by downsampling the majority categories to mitigate class imbalance. The sampling procedures and statistics after sampling are provided in Appendix C.

4.2 Models

Given the multimodal nature of our classification task, involving text from captions, audio transcripts, images, and video content, we require SOTA models designed for multimodal data, particularly video understanding. We evaluate three commercial API-based VLMs: GPT-4.1 (OpenAI, 2025) and Claude-Sonnet-4 (Anthropic PBC, 2025) (both process images), while Gemini-2.5-Flash (Comanici et al., 2025) natively accepts video input. We also evaluate two open-source models: Qwen2.5-VL (Wang et al., 2024), and InternVL3 (Zhu et al., 2025), which uses Qwen2.5 pre-trained base models and Variable Visual Position, which achieves strong performance on video benchmarks.

Model sizes and versions are listed in Table 10, Appendix D.1. Due to computational constraints, we focus on small and medium-sized models. For open-source models, we set *temperature* = 0.1 and *top_p* = 0.9. For API models, we retain the default hyperparameters *temperature* = 1.0 and *top_p* = 1.0. Implementation details and hardware specifications are provided in Appendix D.1. Details about the train and test datasets are in Section C, Appendix.

4.3 Input Features

To capture the full range of pro-bigorexia signals, we extract four complementary input features: **Visual:** For models that natively accept video (e.g., Gemini-2.5-Flash, Qwen2.5VL, InternVL3), we provide raw video. For frame-based models (e.g., Claude-Sonnet-4, GPT-4.1, open-source VLMs), we sample four equally spaced frames per video due to API costs and token-length constraints; ablation studies (Section 4.7) show additional frames yield minimal gains. **Audio:** We apply Google's YAMNet (Hershey et al., 2017) to filter non-speech sounds (music, ambient noise), then transcribe "Speech" segments using OpenAI Whisper (Radford et al., 2022). **On-Screen Text:** We use Gemini-2.5-Flash to extract overlaid text—captions and annotations—since many TikTok creators rely on on-screen text for key messages. **Caption:** We include the original TikTok description (user caption and hashtags) from the video metadata.

4.3.1 Training and Inference Paradigms

Zero-Shot Prompting We create prompts for zero-shot classification by incorporating taxonomy definitions of the primary and subcategories and instructing the model to select the appropriate category as the label. The full zeroshot prompt templates for Tasks 1 and 2 are shown in Figure 6 and 7, Appendix D.2. We evaluate the models on balanced test sets of 588 examples for primary classification (Task 1) and 466 examples for subcategory classification (Task 2), with each class evenly represented (see Appendix C)

Few-Shot Prompting For few-shot experiments, we sample a fixed set of in-context examples from the training data: two videos per primary category (12 examples total) for Task 1 and one video per subcategory (16 examples) for Task 2 (prompt details are in Figures 8 and 9, Appendix). Each example is presented with its full multimodal features: video frames, audio transcript, on-screen text, caption, and the corresponding label. We evaluate the VLMs on a balanced test set of 462 examples (Appendix C).

Finetuning Due to cost constraints, we limit finetuning to open-source VLMs. We instruction-tune various model variants (up to 72B) on our annotated datasets. The batch size is 8, and the learning rate is 5e-5. The inference hyperparameters are similar to zero-shot and few-shot prompting (`temperature` = 0.1 and `top_p` = 0.9).

4.4 Pro-Bigorexia Classification Results

Primary Category Detection Commercial API-based models continue to lead in broad-category detection: Claude-Sonnet-4 (zero-shot) attains the highest accuracy (0.829), macro precision (0.832), macro recall (0.829), and macro F1 (0.827), with its few-shot variant ranking a close second. Although fine-tuned open-source models (e.g., Qwen2.5-VL-32B and InternVL3 variants) gain 10–20 points

Model	Training	Primary Category				Subcategory			
		Acc.	P_m	R_m	$F1_m$	Acc.	P_m	R_m	$F1_m$
GPT-4.1	Zero-shot	0.796	0.808	0.792	0.792	0.652	0.639	0.503	0.532
	Few-shot	0.813	0.820	0.813	0.813	0.680	0.679	0.643	0.639
Claude-Sonnet-4	Zero-shot	**0.829**	**0.832**	**0.829**	**0.827**	0.670	0.679	0.636	0.640
	Few-shot	**0.819**	**0.829**	**0.818**	**0.818**	0.665	0.674	0.627	0.632
Gemini-2.5-Flash	Zero-shot	0.776	0.782	0.776	0.775	0.666	0.665	0.616	0.621
	Few-shot	0.776	0.782	0.776	0.775	0.663	0.645	0.613	0.621
Qwen2.5-VL-7B	Zero-shot	0.539	0.744	0.539	0.526	0.383	0.554	0.327	0.306
	Few-shot	0.689	0.733	0.689	0.689	0.236	0.149	0.142	0.140
	Finetuning	0.784	0.810	0.784	0.776	**0.684**	**0.687**	**0.669**	**0.667**
Qwen2.5-VL-32B	Zero-shot	0.733	0.788	0.733	0.742	0.556	0.637	0.495	0.513
	Few-shot	0.804	0.812	0.804	0.807	0.654	0.662	0.594	0.607
	Finetuning	0.776	0.800	0.776	0.775	0.658	0.662	0.647	0.645
InternVL3-8B	Zero-shot	0.614	0.734	0.614	0.635	0.517	0.582	0.464	0.476
	Few-shot	0.690	0.717	0.690	0.692	0.519	0.560	0.436	0.443
	Finetuning	0.765	0.789	0.765	0.758	**0.690**	**0.686**	**0.679**	**0.675**
InternVL3-38B	Zero-shot	0.784	0.797	0.784	0.785	0.649	0.665	0.603	0.608
	Few-shot	0.806	0.817	0.806	0.807	0.673	0.666	0.630	0.630
	Finetuning	0.767	0.795	0.767	0.758	0.636	0.540	0.535	0.528

Table 2: Classification Macro Metrics for Primary Category and Subcategory by Model and Training Method. Highest values in each column are in **green**, second-highest in **cyan**. P_m, R_m, and $F1_m$ refer to macro precision, macro recall, and macro F1 scores.

over their zero-shot baselines, none yet surpass the commercial zero-shot benchmark. These results confirm that large-scale pretraining still confers an advantage for general category classification, even as domain-specific tuning significantly narrows the gap.

Subcategory Detection In contrast, fine-tuning proves decisive for fine-grained subcategory tasks. InternVL3-8B (finetuned) achieves the top subcategory accuracy (0.690), recall (0.679), and F1 (0.675), while Qwen2.5-VL-7B (finetuned) leads in precision (0.687) and ranks second in other metrics. Commercial zero- and few-shot methods plateau around 0.670 accuracy and underperform relative to these tuned open-source variants. This demonstrates that parameter adaptation is essential to capture the nuanced distinctions of harmful content subtypes.

Prompting vs. Parameter-Tuning Few-shot prompting yields modest boosts in primary-category performance—for instance, GPT-4.1 improves from 0.796 to 0.813 accuracy—highlighting the continued value of emergent prompting abilities. However, in subcategory detection, few-shot gains are minimal, and zero-shot performance remains low. By contrast, parameter-efficient fine-tuning delivers substantial improvements across both tasks, underscoring that direct model adaptation is the most effective strategy for specialized harmful content detection, particularly for granular subcategory classification.

Model Scale and Architecture Larger Qwen2.5-VL-32B boosts zero/few-shot primary category performance, but mid-scale InternVL3-8B (built on the Qwen backbone) achieves the strongest fine-tuning gains—surpassing its 38B variant on subcategory tasks. This underscores that, beyond raw parameter count, architecture critically shapes fine-grained detection performance.

4.5 Severity Score Prediction

Model	Severity	
	MAE ↓	ρ ↑
Claude-Sonnet-4	0.679	**0.675**
GPT-4.1	**0.690**	**0.690**
Gemini-2.5-Flash	**0.690**	**0.690**
InternVL3-8B	0.805	0.474
InternVL3-38B	0.701	0.607
Qwen2.5-VL-7B	0.794	0.484
Qwen2.5-VL-32B	0.688	0.601
Qwen2.5-VL-72B	**0.653**	**0.638**

Table 3: Zero-shot Severity Prediction Metrics by Model (highest in **green**, 2nd highest in **blue**).

To assess the ability of VLMs to predict harm severity, we conducted zero-shot prompting experiments across multiple models. Results reveal that models typically misestimate severity by approximately half a scale point, indicating reasonable but imperfect alignment with expert clinical judgment. Notably, severity prediction proves more challenging than categorical classification, with all models showing only moderate correlation with ground truth ratings. While these findings demonstrate baseline competence in severity estimation, substantial improvement is needed for reliable moderation deployment.

4.6 Error Analysis

Cross-model error analysis exposes fundamental challenges in characterizing pro-bigorexia content. For instance, models consistently misclassify exercise-related content as “Irrelevant” (6.1–14.3%) or confuse “Relationship to Body” with “Relationship to Exercise” (10.2–17.3%), mirroring annotator disagreements (Figures 3 and 4, Appendix B.5). Videos demonstrating workout sessions simultaneously touch on exercise, body display, dieting, lifestyle, and motivations that single-label classification cannot capture.

Classification consistently confuses “Hormone Therapy” and “Anabolic Steroids” (11.8–17.6% bidirectional misclassification), a clinically critical distinction. Models also frequently misclassify “Supplement Abuse” as “Relationship to Body” (8.2–22.4%), suggesting difficulty in recognizing harmful supplement messaging. This likely stems from creators using coded language (e.g., “tren”, “stack”), requiring domain expertise that general-purpose models lack. Additionally, models show systematic bias toward predicting “Muscularity Self-objectification” in subtype classification, with excessive exercise and toxic motivation content frequently misclassified into this category (14.7–47.1%). These systematic errors highlight a fundamental challenge: effective pro-bigorexia detection demands not just multimodal capabilities, but clinical and social knowledge to navigate the blurred boundaries between fitness content and promoting unhealthy behaviors.

4.7 Ablation Study

4.7.1 Input Features: Text vs Video

Table 4 reports results of ablation study to assess the contribution of each modality to classification performance on Task 1 with zero-shot prompting.

Model	Modality	P_m	R_m	$F1_m$
Gemini-2.5-Flash	Audio	0.700	0.423	0.424
	Caption	0.732	0.707	0.709
	OCR	0.717	0.661	0.664
	Text	0.733	0.719	0.719
	Video	**0.786**	**0.765**	**0.766**
	All	**0.782**	**0.776**	**0.775**
GPT-4.1	Audio	0.752	0.444	0.453
	Caption	0.728	0.672	0.679
	OCR	0.741	0.663	0.672
	Text	0.749	0.675	0.681
	Video	**0.755**	**0.697**	**0.701**
	All	**0.808**	**0.792**	**0.792**
InternVL3-38B	Audio	0.632	0.405	0.416
	Caption	0.669	0.590	0.587
	OCR	0.649	0.614	0.618
	Text	0.691	0.630	0.635
	Video	**0.697**	**0.655**	**0.660**
	All	**0.797**	**0.784**	**0.785**

Table 4: Ablation results for task 1 using a single modality input features (text or video). *Audio* modality refers to the transcript of the video; *Caption* refers to the video description; OCR is text within images, and *Text* modality refers to *Caption* + *OCR*. Highest values in each column are in **green**, second-best in **cyan**. P_m, R_m, and $F1_m$ refer to macro precision, macro recall, and macro F1 scores.

We evaluate GPT-4.1 and InternVL3-38B (image-based) and Gemini-2.5-Flash (video-native). Multimodal fusion yields the best results, with GPT-4.1 and InternVL3-38B achieving F1 scores of 0.792 and 0.785, respectively. Video alone provides the strongest unimodal performance (F1 = 0.660–0.701), highlighting the discriminative power of visual behavioral cues. In contrast, audio consistently underperforms (F1 $<$ 0.453), likely due to background music and variable audio quality. Gemini-2.5-Flash shows narrower modality gaps, suggesting native video processing reduces the need for explicit fusion. Overall, results underscore the value of combining visual and textual inputs for robust pro-bigorexia detection.

4.7.2 Number of Frames per Video

We examine whether providing more video content as input features to the models helps improve results. Our frame-count ablation reveals contrasting patterns between models. Performance of GPT-4.1 shows substantial degradation with increased frames (F1: 0.827→0.792→0.685), while InternVL3-38B maintains stable performance across all densities (F1: 0.785-0.790). This suggests that GPT-4.1 suffers from information overload when processing dense temporal se-

Model	Frames #	P_m	R_m	$F1_m$
GPT-4.1	4	0.832	0.829	0.827
	16	0.817	0.793	0.792
	32	0.739	0.657	0.685
InternVL3-38B	4	0.7967	0.7840	0.7849
	16	0.7988	0.7891	0.7896
	32	0.7922	0.7840	0.7844

Table 5: Frame-count ablation for GPT-4.1 and InternVL3-38B. P_m, R_m, and $F1_m$ are macro precision, macro recall, and macro F1.

quences, whereas InternVL3-38B effectively captures the relatively static visual elements in TikTok pro-bigorexia content. The stability across frame counts indicates that complementary text and audio modalities provide sufficient dynamic contextual information, validating our cost-efficient 4-frame approach.

5 Conclusion

We present BIGTOK and BIGTOKDETECT, a clinically-informed dataset and vision–language framework for identifying pro-bigorexia content on TikTok. By curating over 2,200 expert-annotated videos across five primary categories and fine-grained subcategories, we enable systematic study of muscle dysmorphic signals that evade traditional text-based moderation. Our fine-tuned VLMs achieve 0.829% accuracy on primary category classification and 0.690% on subcategory detection, and ablation studies confirm that multimodal fusion of visual, audio, and textual inputs yields a 5–10% boost over unimodal baselines.

Despite these advances, challenges remain in scaling to continually evolving platforms and detecting emerging coded language. Future work expands BIGTOK to include cross-platform content (e.g., Instagram, YouTube Shorts), incorporates temporal dynamics of video sequences, and integrates user-level metadata for context-aware moderation. We also collaborate with clinical practitioners to validate model outputs in real-world settings and explore fairness considerations across demographic groups. By releasing our dataset, code, and model checkpoints, we aim to catalyze further research on robust, ethically grounded multimodal moderation in specialized mental health domains.

Limitations

Dataset and Sampling Constraints Our dataset is limited to English-only TikTok content from 2019-2025, potentially missing cultural variations and platform-specific differences. We focused on English to ensure annotation quality and TikTok as the dominant youth platform. Keyword-based sampling may overlook subtle or emerging pro-bigorexia content that avoids our taxonomy terms, though we exhaustively developed keywords through previous literature review and expert consultation to maximize coverage.

Annotation and Taxonomy Limitations Forced reduction from multi-faceted expert annotations to single labels loses nuanced information. We simplified to single labels due to current LLM limitations in reliable multi-label classification. Our 16 annotators are predominantly female (13/16), potentially introducing gender perspective bias in male-centric bigorexia evaluation; however, this field remains understudied, and we could not find sufficient male experts specializing in bigorexia. Our taxonomy represents a living document that may miss emerging patterns, though we strived for exhaustiveness through iterative expert refinement.

Model Selection Limitations While we selected current leading VLMs across both commercial and open-source categories, the rapidly evolving landscape means we may be missing other capable models. We prioritized models with proven multimodal video understanding capabilities, but acknowledge that newer or specialized architectures might offer different performance characteristics.

Experimental Reproducibility Due to computational cost constraints, we conduct single runs for each experimental configuration without multiple trials or statistical significance testing. While our results establish baseline performance across models and tasks, future work should include multiple experimental runs with statistical analysis to provide more robust performance comparisons and confidence intervals for the reported metrics.

Ethics Statement

This research addresses harmful content classification, which raises important ethical considerations. To protect annotators' well-being when reviewing potentially disturbing pro-bigorexia content, we exclusively recruited clinical experts—licensed psychologists, psychiatrists, and doctoral candidates specializing in eating disorders—who possess the professional training necessary to safely engage with this sensitive material. Their clinical expertise provides essential psychological safeguards while ensuring high-quality annotations. While our work aims to support mental health research, we acknowledge that automated detection systems risk censorship or discrimination against legitimate fitness communities. We took precautions, including content anonymization and clinical expert consultation, focusing on detection research rather than deployment recommendations. **Our work is intended for research purposes and should not be used for clinical diagnosis of bigorexia without proper clinical oversight.** All procedures were conducted in accordance with institutional review board guidelines.

We acknowledge the use of AI language models to assist with writing and editing portions of this manuscript. All AI-generated content was reviewed, verified, and edited by the authors. The research design, data collection, analysis, and primary contributions remain entirely the work of the human authors.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, and 8 others. 2022. Flamingo: a visual language model for few-shot learning. *Preprint*, arXiv:2204.14198.

Anthropic PBC. 2025. System card: Claude opus 4 & claude sonnet 4. Technical report, Anthropic PBC. Accessed: 2025-07-27.

Corey H Basch, Lorie Donelle, Joseph Fera, and Christie Jaime. 2022. Deconstructing tiktok videos on mental health: Cross-sectional, descriptive content analysis. *JMIR Form Res*, 6(5):e38340.

Anne E Becker. 2004. Television, disordered eating, and young women in fiji: Negotiating body image and identity during rapid social change. *Culture, medicine and psychiatry*, 28:533–559.

Charles Bickham, Bryan Ramirez-Gonzalez, Minh Duc Chu, Kristina Lerman, and Emilio Ferrara. 2025. Edtok: A dataset for eating disorder content on tiktok. *Preprint*, arXiv:2505.02250.

Stevie Chancellor and Munmun Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *NPJ Digital Medicine*, 3(1):43.

Stevie Chancellor, Yannis Kalantidis, Jessica A. Pater, Munmun De Choudhury, and David A. Shamma. 2017. Multimodal classification of moderated online pro-eating disorder content. In *Proceedings of the 2017 ACM CHI Conference on Human Factors in Computing Systems*, pages 3213–3226. ACM.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, Luke Marris, Sam Petulla, Colin Gaffney, Asaf Aharoni, Nathan Lintz, Tiago Cardal Pais, Henrik Jacobsson, Idan Szpektor, Nan-Jiang Jiang, and 3290 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *Preprint*, arXiv:2507.06261.

Marita Cooper, Kamryn T. Eddy, Jennifer J. Thomas, Debra L. Franko, Bradley Carron-Arthur, Ani C. Keshishian, and Kathleen M. Griffiths. 2020. Muscle dysmorphia: A systematic and meta-analytic review of the literature to assess diagnostic validity. *International Journal of Eating Disorders*, 53(10):1583–1604.

Melissa Donati, Ludovica Polidori, Paola Vernillo, and Gloria Gagliardi. 2023. Building a corpus on eating disorders from TikTok: Challenges and opportunities. In *Proceedings of the 9th Italian Conference on Computational Linguistics (CLiC-it 2023)*, pages 515–520, Venice, Italy. CEUR Workshop Proceedings.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *Preprint*, arXiv:2405.21075.

Kyle T. Ganson, Alexander Testa, Rachel F. Rodgers, and Jason M. Nagata. 2025. Associations between muscularity-oriented social media content and muscle dysmorphia among boys and men. *Body Image*, 53:101903.

Tarleton Gillespie. 2018. *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media.*

David Gimeno-Gómez, Ana-Maria Bucur, Adrian Cosma, Carlos-David Martínez-Hinarejos, and Paolo Rosso. 2024. Reading between the frames: Multimodal depression detection in videos from non-verbal cues. *Preprint*, arXiv:2401.02746.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. 2020. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7:205395171989794.

Shawn Hershey, Sourish Chaudhuri, Daniel P. W. Ellis, Jort F. Gemmeke, Aren Jansen, R. Channing Moore, Manoj Plakal, Devin Platt, Rif A. Saurous, Bryan Seybold, Malcolm Slaney, Ron J. Weiss, and Kevin Wilson. 2017. Cnn architectures for large-scale audio classification. *Preprint*, arXiv:1609.09430.

Nick Ara Kamkari. 2025. The manosphere and men's health: unpacking the links between online communities, body dysmorphia and erectile dysfunction. *Journal of Men's Health*, 21(4):123–125.

Douwe Kiela, Suvrat Bhooshan, Hamed Firooz, Ethan Perez, and Davide Testuggine. 2020a. Supervised multimodal bitransformers for classifying images and text. *Preprint*, arXiv:1909.02950.

Douwe Kiela, Hamed Firooz, Aravind Mohan, Vedanuj Goswami, Amanpreet Singh, Pratik Ringshia, and Davide Testuggine. 2020b. 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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *Preprint*, arXiv:2301.12597.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Tackling multi-image, video, and 3d in large multimodal models. In *arXiv preprint arXiv:2407.07895*.

Valerie Lookingbill, Ehsan Mohammadi, and Yizhou Cai. 2023. Assessment of accuracy, user engagement, and themes of eating disorder content in social media short videos. *JAMA Network Open*, 6(4):e238897.

Marcus Ma, Georgios Chochlakis, Niyantha Maruthu Pandiyan, Jesse Thomason, and Shrikanth Narayanan. 2025. Large language models do multi-label classification differently. *Preprint*, arXiv:2505.17510.

Ghofrane Merhbene, Alexandre Puttick, and Mascha Kurpicz-Briki. 2024. Investigating machine learning and natural language processing techniques applied for detecting eating disorders: a systematic literature review. *Frontiers in Psychiatry*, 15:1319522.

Maria Minadeo and Lizzy Pope. 2022. Weight-normative messaging predominates on tiktok—a qualitative content analysis. *PLOS ONE*, 17(11):e0267997.

Deborah Mitchison, Jonathan Mond, Scott Griffiths, Phillipa Hay, Jason M. Nagata, Kay Bussey, Nora Trompeter, Alexandra Lonergan, and Stuart B. Murray. 2022. Prevalence of muscle dysmorphia in adolescents: findings from the everybody study. *Psychological Medicine*, 52(14):3142–3149.

Stuart Murray, Leila Hazery, Tori Shen, Tom Wooldridge, and Jonathan Mond. 2015. Go big or go home: A thematic content analysis of promuscularity websites. *Body image*, 16:17–20.

Stuart B Murray, Jason M Nagata, Scott Griffiths, Jerel P Calzo, Tiffany A Brown, Deborah Mitchison, Aaron J Blashill, and Jonathan M Mond. 2017. The enigma of male eating disorders: A critical review and synthesis. *Clinical psychology review*, 57:1–11.

OpenAI. 2023. Gpt-4 technical report. Online. `https://openai.com/research/gpt-4`.

OpenAI. 2025. Gpt-4.1. `https://openai.com/index/gpt-4-1/`. Accessed: 2025-07-27.

Harrison G. Pope Jr, Anthony J. Gruber, Paulette Choi, Roberto Olivardia, and Katharine A. Phillips. 1997. Muscle dysmorphia: An underrecognized form of body dysmorphic disorder. *Psychosomatics*, 38(6):548–557.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

TikTok. 2025. Query videos – tiktok research api. Online; accessed July 2025.

TikTok Creative Center. Top hashtags overview. `https://ads.tiktok.com/business/creativecenter/hashtag/pc/en`. Accessed January 2025.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Tao Wang, Markus Brede, Antonella Ianni, and Emmanouil Mentzakis. 2017. Detecting and characterizing eating-disorder communities on social media. In *Proceedings of the Tenth ACM International Conference on Web Search and Data Mining*, pages 91–100.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *Preprint*, arXiv:2403.13372.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, Zhangwei Gao, Erfei Cui, Yinan Liu, Tan Jiang, Jiapeng Luo, Yi Wang, Yinan He, Tan Lu, Lewei Lu, and 4 others. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Appendix

A Taxonomy

A.1 Categories Definition

A detailed definition of the categories in our taxonomy is provided in Table 12. These categories are carefully curated and iteratively refined in collaboration with clinical psychologists and domain researchers to ensure both validity and practicality.

B Annotation

B.1 Annotator Profiles

We enlist 16 subject-matter experts, ranging from clinical psychologists and psychiatrists to computational social scientists, for video annotation. Our annotators are research collaborators and co-authors who volunteered their clinical expertise with full knowledge of the project's scope and scientific objectives, ensuring informed participation in this sensitive content annotation task. Table 11 details each annotator's ID, batch assignment, area of expertise, and gender.

B.2 Annotation Instructions

Annotators received comprehensive guidelines emphasizing safety protocols and decision-making frameworks for handling potentially disturbing content. Key instructions included: (1) **Content Safety**: annotators were advised to work at their own pace, take regular breaks, and prioritize mental well-being when reviewing harmful content; (2) **Decision Framework**: when encountering ambiguous cases, annotators were instructed to select "Unsure, need consultation" and provide explanatory notes, with textual information (captions, on-screen text) prioritized over conflicting visual content; (3) **Consistency Maintenance**: annotators were asked to reflect on their annotation patterns periodically to avoid drift and maintain consistency across the 300-video batches; (4) **Expert Support**: availability of research coordinators via dedicated communication channels for immediate consultation on challenging cases.

Annotators assess harm severity based on message intensity, including explicit display of muscular physiques, toxic motivational pressure toward unattainable muscularity, rigid and unrealistic dietary demands, endorsement of high-risk behaviors such as steroid use or dangerous workout practices, and the degree of pathological behavioral patterns exhibited in the content.

These protocols ensured both annotator safety and annotation quality while acknowledging the inherently difficult nature of distinguishing harmful pro-bigorexia content from benign fitness material.

B.3 Annotation Platform

The HTML interface (Figure 2) presents each TikTok video with its caption and hashtags (if any). The form beneath the video prompts the annotator to mark if the video is: (1) *Showing One's Body* (yes/no); includes (2) *Commercial Sponsorship* (yes/no); (3) 2 boxes for *Type of harm* (1 mandatory, 1 optional), consisting of primary categories and subcategories, plus "Relevant but Not Listed," "Irrelevant," and "Unsure—Needs Consultation" options; and (4) *Severity of Harm* on a discrete scale ranging from 1 (Not Harmful) to 5 (Extremely Harmful). A text box allows annotators to write additional comments and edge-case observations.

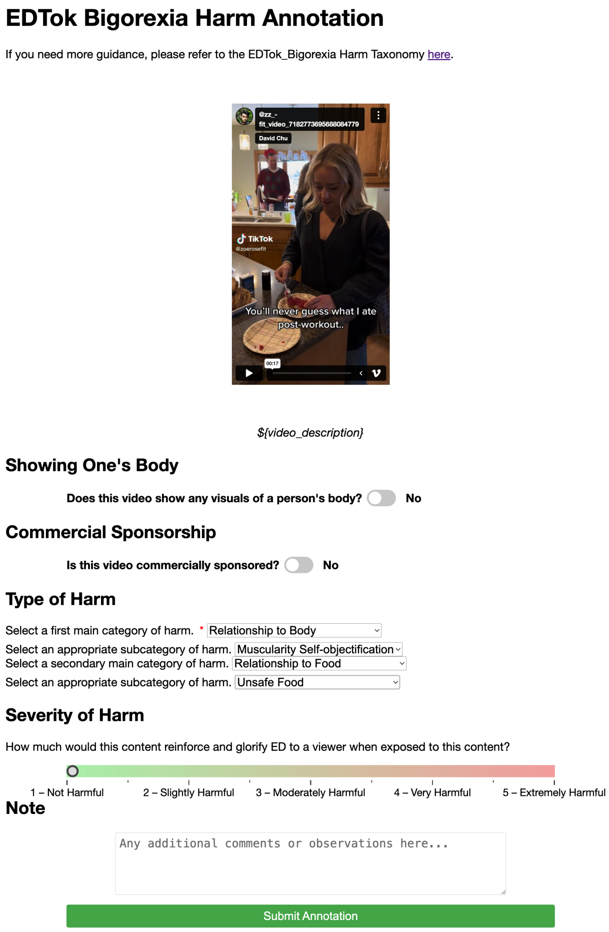

Figure 2: Annotation Interface on Amazon Mechanical Turk platform.

B.4 Annotation Processing

After Round 2, we apply a five-step pipeline to prepare the data for consensus review and final curation:

1. **Annotation Completeness:** verify that each video has exactly two independent annotations.
2. **Data Filtering:** remove videos that are unavailable (deleted) or non-English, as indicated in annotators' notes.
3. **Reliability Assessment:** compute Cohen's κ for both strong (exact primary–secondary match) and weak (primary-only match) agreement to quantify inter-rater reliability.
4. **Category Refinement Candidates:** collect all videos flagged as "Unsure—Needs Consultation" or "Relevant but Not Listed." For these, we consult with the annotators to either reassign them to existing categories, extend the taxonomy to cover frequently emerging content types, or mark them as irrelevant.
5. **Consensus Review and Re-annotation:** filter out all items by agreement level (strong, weak, disagreement), sample a subset of weak and disagreement cases for each annotator pair, hold a batch-specific meeting to discuss and jointly re-annotate selected cases to reach consensus, then return the remaining weak and disagreement items to annotators for independent re-annotation.

This pipeline ensures data consistency, highlights contentious cases for Round 3 consensus sessions, and informs any necessary taxonomy updates.

B.5 Annotation Statistics

Inter-annotator agreement substantially improves from Round 2 to Final, with Cohen's κ values increasing from moderate agreement (0.43-0.69 strong κ, 0.59-0.83 weak κ) to good-to-excellent agreement (0.58-0.81 strong κ, 0.78-0.94 weak κ). The consistently high weak κ values (>0.9 in most batches) in the Final round indicate that annotators achieved near-perfect agreement on the broader annotation categories.

C Data

We collected videos through the official TikTok Research API after submitting a research proposal that was reviewed and approved by TikTok. This API accesses publicly posted content where users

Batch	Round 2		Final	
	Strong k	Weak k	Strong k	Weak k
1	0.551	0.613	0.616	0.775
2	0.570	0.588	0.762	0.841
3	0.510	0.612	0.680	0.815
4	0.551	0.785	0.811	0.937
5	0.563	0.685	0.787	0.924
6	0.555	0.723	0.742	0.938
7	0.430	0.739	0.577	0.931
8	0.685	0.825	0.769	0.938

Table 6: Cohen's kappa coefficients for the inter-agreement rate among annotators in Rounds 2 and 3

Primary Category	**Count**
Relationship to Body	1442
Relationship to Exercise	1233
Relationship to Food	824
Supplement Abuse	558
Relevant but Not Harmful	411
Relationship to Masculinity	406
Irrelevant	401

(a) Primary Categories

Subcategory	**Count**
Muscularity Self-objectification	905
Leanness Self-objectification	238
Muscle Dissatisfaction	240
Rigid Food Rules	375
Cheat Meals	193
Excessive Exercise	379
Maladaptive Coping	213
Toxic Motivation	308
Anabolic Steroids	195
Other (please specify in Note)	217

(b) Subcategories (reordered)

Table 7: Distribution of Annotated Primary Categories and Subcategories

have consented to public visibility through TikTok's terms of service. We implemented additional privacy protections by anonymizing user identifiers and exposing only video content and captions to annotators.

Our dataset comprises TikTok videos annotated for hierarchical classification across two levels of granularity. The primary category classification categorizes content into six broad categories related to body image and health behaviors, while the subcategory classification provides fine-grained classification into 16 specific subcategories. Severity score scale define the intensity of harm emotion. The severity scores are most heavily concentrated in the 1.0 to 1.5 range (Figure 5), indicating that lower severity levels are the most common. Frequency then gradually decreases as scores increase, with relatively few cases above 4.0. This suggests that mild severity is predominant in the dataset. We remove corrupted files (e.g., missing metadata) and non-English videos during preprocessing. The dataset is split into train/test sets, maintaining approximately a 3:1 ratio with stratified sampling based on task labels and downsampling of dominant categories to improve balance.

As shown in Table 8, the primary category task contains 1,966 training and 588 test samples. Training data exhibits natural class imbalance reflecting real-world distribution, ranging from 489 samples ("Relationship to Body") to 66 samples ("Relationship to Masculinity"), while the test set maintains balanced representation (98 samples per category) for fair evaluation. The subcategory task operates on a filtered subset of 1,472 training and 462 test samples (Table 9). Training samples range from 200 ("Irrelevant") to 28 ("Predebting Exercise"), while test data maintains balanced distribution for major categories (34 samples each) with reduced representation for rare categories (4-10 samples).

Zero-shot and few-shot approaches are evaluated solely on test sets, while finetuning models are trained on the respective task's training set and evaluated on the corresponding test set.

Primary Category	**Train**	**Test**
Relationship to Body	489	98
Irrelevant	450	98
Relationship to Exercise	450	98
Relationship to Food	310	98
Supplement Abuse	201	98
Relationship to Masculinity	66	98
Total	**1,966**	**588**

Table 8: Distribution of samples across primary categories (Task 1).

D Modeling

D.1 Experiment Setup

We implement the VLMs using Transformers-based (for InternVL3) and vLLM (Kwon et al., 2023) (for Qwen-2.5VL) implementations through the LlamaFactory framework (Zheng et al., 2024), which provides efficient inference and serving capabilities for VLMs. We used pre-trained VLMs according to their respective licenses and terms of use: commercial API-based models (GPT-4.1, Claude-Sonnet-4, Gemini-2.5-Flash) under their standard API terms, and open-source models (InternVL3,

Subcategory	Train	Test
Muscularity Self-objectification	170	34
Leanness Self-objectification	59	34
Muscle Dissatisfaction	72	34
Rigid Food Rules	144	34
Unsafe Food	59	34
Cheat Meals	60	34
Excessive Exercise	129	34
Predebting Exercise	28	4
Maladaptive Coping	82	34
Exercise-Induced Functional Impairment	43	6
Toxic Motivation	102	34
Anabolic Steroids	61	34
Legal APEDs	56	10
Hormone Therapy	77	34
Relationship to Masculinity	130	34
Irrelevant	200	34
Total	**1,472**	**462**

Table 9: Distribution of samples across subcategories (Task 2).

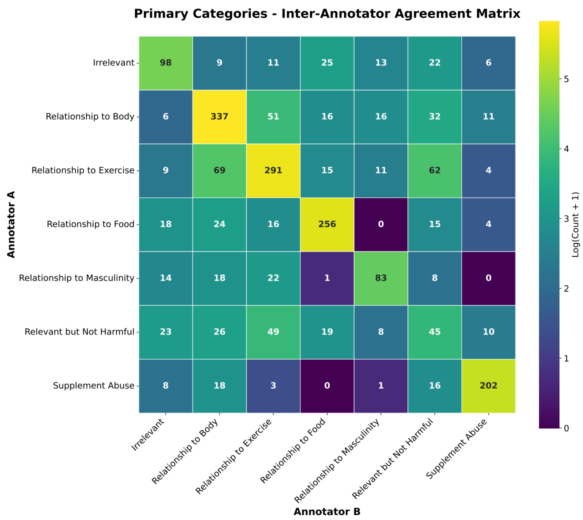

Figure 3: Inter-annotator agreement matrix for primary categories. Each cell shows the number of videos where Annotator A assigned the row category and Annotator B assigned the column category. Diagonal elements represent perfect agreement, while off-diagonal elements indicate disagreements between annotators.

Qwen2.5-VL) under their permissive licenses for research use.

Due to hardware constraints, we exclude InternVL-38B and InternVL-78VL from certain experiments. Our current infrastructure is incompatible with the vLLM version required for multi-node deployment of these larger models. Since these models exceed the memory capacity available on single nodes in our cluster, we cannot accommodate their substantial memory requirements within our computational environment.

We consider conducting experiments with LLaVA-NeXT-Video (Liu et al., 2024), which achieved SOTA open-source performance on Video-MME (Fu et al., 2025). However, we exclude LLaVA-NeXT-Video from our study since its context window limitation (4,096 tokens) cannot accommodate our prompting setup (text + video/frames), which often exceeds that limit.

All zero-shot experiments on open-source VLMs are conducted on 8 $\times$ NVIDIA H100 GPUs. For few-shot and finetuning experiments, we use 32 $\times$ NVIDIA A100 GPUs for Qwen2.5-VL-7B, Qwen2.5-VL-32B, InternVL3-8B, and InternVL3-38B, and 64 $\times$ NVIDIA A100 GPUs for Qwen2.5-VL-72B.

We focus on evaluating two tasks: primary category and subcategory classification. For each task, we split the data into training and test sets at a 3:1 ratio using stratified sampling. The test set is strictly balanced across all classes, while the training set is adjusted by downsampling the majority categories to mitigate class imbalance. Detailed split statistics and sampling procedures are provided in Appendix C.

Model	Version	Size (B)	Type
GPT-4o	2024-08-06	—	API
Claude Sonnet 4	2024-02-24	—	API
Gemini 2.5 Pro	2025-06-17	—	API
InternVL3	2024-10-04	8, 38	Open
Qwen2.5-VL	2025-06-05	7, 32, 72	Open

Table 10: Model version dates, parameter counts, and types (API vs. Open-source) for video-based VLMs.

D.2 Prompt Templates

We use a JSON-based prompt structure following the standard role-content format for chat-based LLMs called 'sharegpt'. Each prompt consists of:

- A sequence of images from the video (sampled frames)
- A user message containing the video caption, embedded text and audio transcription
- A single assistant response with the predicted label

An example format of prompt used for zero-shot inference/supervised finetuning and few-shot inference is shown in Figures 6 and 7 and Figure 8 and 9:

ID	Batch	Expertise	Gender
A1	1	clinical psychology doctoral candidate specializing in male eating disorders and muscle dysmorphia intervention research.	Male
A2	1	senior researcher in computational social science focused on social media dynamics and their implications for mental health.	Female
A3	2	health communication professor using computational social network methods to study online health behaviors.	Female
A4	2	medical student focusing on clinical and translational research in body image disorders and digital health interventions.	Female
A5	3	clinical psychology doctoral candidate specializing in digital mental health and disordered eating prevention.	Female
A6	3	communication doctoral candidate investigating how interactive media–driven social comparisons affect body image outcomes.	Female
A7	4	clinical professor and psychiatrist studying the psychopathology and treatment of eating disorders and muscle dysmorphia.	Male
A8	4	clinical professor and pediatric psychologist specializing in adolescent body image disorders.	Female
A9	5	child and adolescent psychiatrist specializing in developmental psychopathology and body image disorders.	Female
A10	5	clinical social worker specializing in child and adolescent body image disorders and psychosocial interventions in neuropsychiatric care.	Female
A11	6	occupational therapist specializing in functional rehabilitation and psychosocial support for eating-disorder patients in a neuropsychiatric hospital setting.	Female
A12	6	clinical dietitian specializing in nutritional assessment and dietary management for eating and body-image disorders.	Female
A13	7	mental health nurse practitioner specializing in the assessment and treatment of child and adolescent psychiatric disorders.	Female
A14	7	board-certified child and adolescent psychiatrist specializing in mood and behavioral disorders.	Female
A15	8	clinical social worker with inpatient psychiatric experience in providing psychotherapy to adult patients.	Female
A16	8	doctoral candidate specializing in detecting disordered behaviors in social media.	Male

Table 11: Annotator assignments, areas of expertise, and gender.

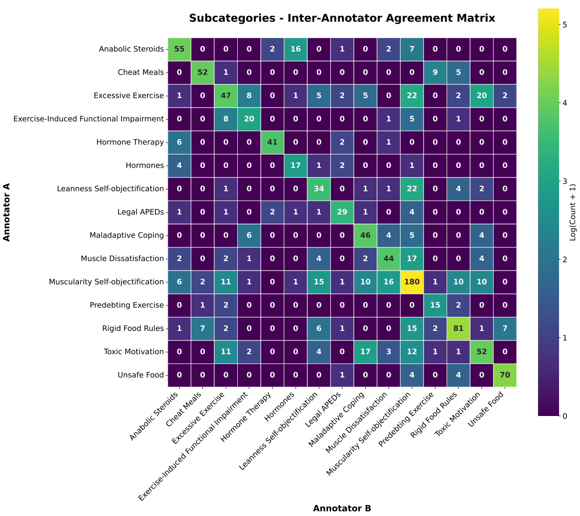

Figure 4: Inter-annotator agreement matrix for subcategories. Each cell shows the number of videos where Annotator A assigned the row subcategory and Annotator B assigned the column subcategory. The matrix reveals patterns of confusion between semantically related subcategories and overall annotation consistency.

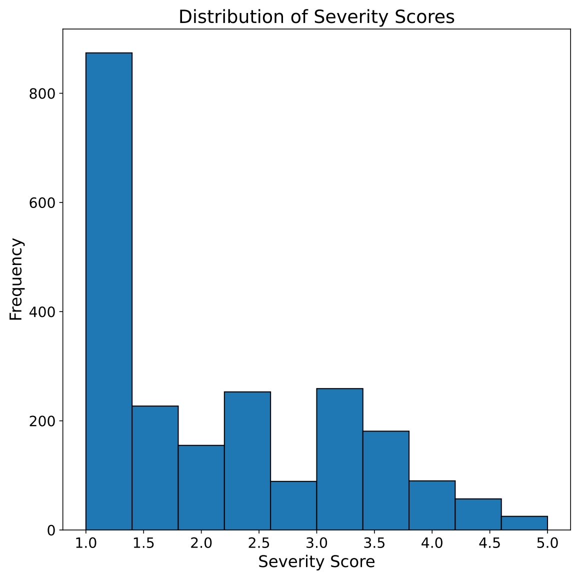

Figure 5: Distribution of Severity Scores. The histogram shows most values concentrated between 1.0 and 2.0, with fewer cases at higher severity levels.

VLM Zero-Shot/SFT Prompt - Task 1: Primary Category Classification

```
<video>You are an expert classifier of harmful pro-muscle-dysmorphia content
on TikTok.
Taxonomy:
Type 1: Relationship to Body:
  • Subtype 1: Muscularity Self-Objectification:Idealized imagery emphasizing
  muscular aesthetics.
  • Subtype 2: Leanness Self-Objectification: Self-evaluation against a lean ideal.
  • Subtype 3: Muscle Dissatisfaction: Perceived insufficient muscularity.

Type 2: Relationship to Food:
  • Subtype 1: Rigid Food Rules: Obsessive tracking, bulking, or cutting.
  • Subtype 2: Unsafe Food: Promoting unconventional muscle-enhancing foods.
  • Subtype 3: Cheat Meals: Binge-compensation cycles.

Type 3: Supplement Abuse:
  • Subtype 1: Anabolic Steroids: Downplaying risks of steroid use.
  • Subtype 2: Legal APEDs: Excessive legal supplement usage.
  • Subtype 3: Hormone Therapy: Promoting TRT as performance enhancer.

Type 4: Relationship to Exercise:
  • Subtype 1: Excessive Exercise despite harm.
  • Subtype 2: Predebting: Exercising to justify eating.
  • Subtype 3: Maladaptive Coping: Sole coping method.
  • Subtype 4: Functional Impairment.
  • Subtype 5: Toxic Motivation.

Type 5: Relationship to Masculinity:
  • Subtype 1: Linking muscles to male identity and worth.

Type 6: Irrelevant:
  • Subtype 1: Content unrelated to the above.

Carefully consider all sources of information about this video:

VIDEO_ID: <Anonymized Video ID>
Caption: <TikTok Video Caption>
Audio transcript: <Audio Transcriptions using Whisper>
Embedded text: <OCR text from video>

###TASK###
Classify the video into one of the following types:
Relationship to Body, Relationship to Food, Supplement Abuse,
Relationship to Exercise, Relationship to Masculinity, or
Irrelevant.

Only output the type label, no explanations or subtypes.
```

Figure 6: Prompt used for zero-shot inference and supervised finetuning for Task 1.

VLM Zero-Shot/SFT Prompt - Task 2: Subcategory Classification

```
<video>You are an expert classifier of harmful pro-muscle-dysmorphia content
on TikTok.
Taxonomy:
Type 1: Relationship to Body:
  • Subtype 1: Muscularity Self-Objectification:Idealized imagery emphasizing
  muscular aesthetics.
  • Subtype 2: Leanness Self-Objectification: Self-evaluation against a lean ideal.
  • Subtype 3: Muscle Dissatisfaction: Perceived insufficient muscularity.

Type 2: Relationship to Food:
  • Subtype 1: Rigid Food Rules: Obsessive tracking, bulking, or cutting.
  • Subtype 2: Unsafe Food: Promoting unconventional muscle-enhancing foods.
  • Subtype 3: Cheat Meals: Binge-compensation cycles.

Type 3: Supplement Abuse:
  • Subtype 1: Anabolic Steroids: Downplaying risks of steroid use.
  • Subtype 2: Legal APEDs: Excessive legal supplement usage.
  • Subtype 3: Hormone Therapy: Promoting TRT as performance enhancer.

Type 4: Relationship to Exercise:
  • Subtype 1: Excessive Exercise despite harm.
  • Subtype 2: Predebting: Exercising to justify eating.
  • Subtype 3: Maladaptive Coping: Sole coping method.
  • Subtype 4: Functional Impairment.
  • Subtype 5: Toxic Motivation.

Type 5: Relationship to Masculinity:
  • Subtype 1: Linking muscles to male identity and worth.

Type 6: Irrelevant:
  • Subtype 1: Content unrelated to the above.

Carefully consider all sources of information about this video:

VIDEO_ID: <Anonymized Video ID>
Caption: <TikTok Video Caption>
Audio transcript: <Audio Transcriptions using Whisper>
Embedded text: <OCR text from video>

###TASK###
Classify the video into one of the following subtypes: Muscularity Self-Objectification,
Leanness Self-Objectification,
Muscle Dissatisfaction, Rigid Food Rules,
Unsafe Food, Cheat Meals,
Anabolic Steroids, Legal APEDs,
Hormone Therapy, Excessive Exercise,
Predebting, Maladaptive Coping,
Exercise-Induced Functional Impairment,
Toxic Motivation,
Relationship to Masculinity, or Irrelevant.

Only output the specific subtype label, no explanations or other text.
If none apply, output Irrelevant.
Use the exact subtype names, not the type names.
```

Figure 7: Prompt used for zero-shot inference and supervised finetuning for Task 2.

VLM Few-Shot Prompt. Task 1: Primary Category Classification

```
You are an expert classifier of harmful pro-muscle-dysmorphia content on TikTok.
Taxonomy:
Type 1: Relationship to Body:
  • Subtype 1: Muscularity Self-Objectification: Idealized imagery emphasizing muscular
  aesthetics as the primary source of value.
  • Subtype 2: Leanness Self-Objectification: Self-evaluation against a lean,
  low-fat, highly-toned ideal.
  • Subtype 3: Muscle Dissatisfaction: Expressing perceived insufficient muscularity despite
  having a muscular physique.
Type 2: Relationship to Food:
  • Subtype 1: Rigid Food Rules: Obsessive macro/micronutrient tracking
  and restrictive dietary practices to rapidly gain muscle mass (bulking) or lose fat (cutting).
  • Subtype 2: Unsafe Food: Promotion of unconventional foods believed
  to enhance muscle growth.
  • Subtype 3: Cheat Meals: Large "reward" meals after restrictive dieting
  that reinforce binge-compensation cycles.
Type 3: Supplement Abuse:
  • Subtype 1: Anabolic Steroids: Normalization or
  endorsement of anabolic-androgenic steroid use with downplayed risks.
  • Subtype 2: Legal APEDs: Overuse of legal supplements (e.g., creatine,
  protein, pre-workout) beyond recommended doses.
  • Subtype 3: Hormone Therapy: Downplaying risks and spreading misinformation
  about testosterone replacement therapy, promoted as a performance enhancer or "anti-aging" treatment
  without proper medical diagnosis.
Type 4: Relationship to Exercise:
  • Subtype 1: Excessive Exercise: Extreme exercise routines exceeding healthy limits despite
  injury or life interference.
  • Subtype 2: Predebting: Treating exercise as punishment or permission to eat.
  • Subtype 3: Maladaptive Coping: Using exercise as the sole coping mechanism to avoid
  emotional distress.
  • Subtype 4: Exercise-Induced Functional Impairment: Prioritizing exercise
  over essential duties, harming daily functioning.
  • Subtype 5: Toxic Motivation: Demeaning communication that pressures
  unrealistic fitness standards via shaming or slurs.
Type 5: Relationship to Masculinity:
  • Subtype 1: Linking muscle-building and exercise performance to male identity,
  sexuality, and self-worth.
Type 6: Irrelevant:
  • Subtype 1: Content without muscle-obsession, restrictive
  diets, supplement/AAS promotion, or extreme exercise. Includes dance trends,
  memes, travel/cooking vlogs, general wellness (e.g., yoga), and pure entertainment unrelated
  to body-image narratives.
Here are some examples from the training data:
Example 1:
<video>
Caption: <Video Caption>
Audio transcript: <Audio Transcription from Whisper>
Embedded text: <OCR Text>
Classification: <Type>
.
.
.
Example 12:
<video>
Caption: <Video Caption>
Audio transcript: <Audio Transcription from Whisper>
Embedded text: <OCR Text>
Classification: <Type>
<video>
VIDEO_ID:
Caption:
Audio transcript:
Embedded text:

###TASK###
Classify the video into one of the following types: Relationship to Body, Relationship to Food,
Supplement Abuse, Relationship to Exercise, Relationship to Masculinity, or Irrelevant.
Only output the type label, no explanations or other text. If none apply, output Irrelevant.
Don't use a subtype, such as "Muscularity Self-Objectification" or "Unsafe Food".
Valid outputs: Relationship to Body, Relationship to Food, Supplement Abuse, Relationship to Exercise,
Relationship to Masculinity, Irrelevant
```

Figure 8: Prompt used for few-shot inference for Task 1.

VLM Few-Shot Prompt. Task 2: Subcategory Classification

```
You are an expert classifier of harmful pro-muscle-dysmorphia content on TikTok.
Taxonomy:
Type 1: Relationship to Body:
  • Subtype 1: Muscularity Self-Objectification: Idealized imagery emphasizing muscular
  aesthetics as the primary source of value.
  • Subtype 2: Leanness Self-Objectification: Self-evaluation against a lean,
  low-fat, highly-toned ideal.
  • Subtype 3: Muscle Dissatisfaction: Expressing perceived insufficient muscularity despite
  having a muscular physique.
Type 2: Relationship to Food:
  • Subtype 1: Rigid Food Rules: Obsessive macro/micronutrient tracking
  and restrictive dietary practices to rapidly gain muscle mass (bulking) or lose fat (cutting).
  • Subtype 2: Unsafe Food: Promotion of unconventional foods believed
  to enhance muscle growth.
  • Subtype 3: Cheat Meals: Large "reward" meals after restrictive dieting
  that reinforce binge-compensation cycles.
Type 3: Supplement Abuse:
  • Subtype 1: Anabolic Steroids: Normalization or
  endorsement of anabolic-androgenic steroid use with downplayed risks.
  • Subtype 2: Legal APEDs: Overuse of legal supplements (e.g., creatine,
  protein, pre-workout) beyond recommended doses.
  • Subtype 3: Hormone Therapy: Downplaying risks and spreading misinformation
  about testosterone replacement therapy, promoted as a performance enhancer or "anti-aging" treatment
  without proper medical diagnosis.
Type 4: Relationship to Exercise:
  • Subtype 1: Excessive Exercise: Extreme exercise routines exceeding healthy limits despite
  injury or life interference.
  • Subtype 2: Predebting: Treating exercise as punishment or permission to eat.
  • Subtype 3: Maladaptive Coping: Using exercise as the sole coping mechanism to avoid
  emotional distress.
  • Subtype 4: Exercise-Induced Functional Impairment: Prioritizing exercise
  over essential duties, harming daily functioning.
  • Subtype 5: Toxic Motivation: Demeaning communication that pressures
  unrealistic fitness standards via shaming or slurs.
Type 5: Relationship to Masculinity:
  • Subtype 1: Linking muscle-building and exercise performance to male identity,
  sexuality, and self-worth.
Type 6: Irrelevant:
  • Subtype 1: Content without muscle-obsession, restrictive
  diets, supplement/AAS promotion, or extreme exercise. Includes dance trends,
  memes, travel/cooking vlogs, general wellness (e.g., yoga), and pure entertainment unrelated
  to body-image narratives.
Here are some examples from the training data:
Example 1:
<video>
Caption: <Video Caption>
Audio transcript: <Audio Transcription from Whisper>
Embedded text: <OCR Text>
Classification: <Type>
.
.
.
Example 12:
<video>
Caption: <Video Caption>
Audio transcript: <Audio Transcription from Whisper>
Embedded text: <OCR Text>
Classification: <Type>
<video>
VIDEO_ID:
Caption:
Audio transcript:
Embedded text:

###TASK###
Classify the video into one of the following subtypes: Muscularity Self-Objectification,
Leanness Self-Objectification,
Muscle Dissatisfaction, Rigid Food Rules,
Unsafe Food, Cheat Meals,
Anabolic Steroids, Legal APEDs,
Hormone Therapy, Excessive Exercise,
Predebting, Maladaptive Coping,
Exercise-Induced Functional Impairment,
Toxic Motivation,
Relationship to Masculinity, or Irrelevant.
```

Figure 9: Prompt used for few-shot inference for Task 2.

Primary	**Secondary**	**Definition**	**Keywords**
Relationship to Body	Muscularity Self-Objectification	Idealized imagery emphasizing muscular aesthetics as the primary source of value.	shredded, swole, mensphysique, bodybuilding
	Leanness Self-Objectification	Self-evaluation against a lean, low-fat, highly-toned ideal promoted online.	flat tummy, skinny men physique, small waist fitness
	Muscle Dissatisfaction	Expressing perceived insufficient muscularity despite fitness, fueling negative self-view.	muscles never big enough, muscles not big enough, muscle dysmorphia, bdd men
Relationship to Food	Rigid Food Rules	Obsessive macro/micronutrient tracking and extreme bulking or cutting diets.	aggressive cut, aggressive bulk, shredded diet, macro tracking
	Unsafe Food	Promotion of raw or unsafe foods believed to enhance muscle growth.	liver king diet, raw meat diet for gym, dog food to gain muscles
	Cheat Meals	Large "reward" meals after restrictive dieting that reinforce binge–compensation cycles.	cheatday food, cheatmeal
Relationship to Exercise	Excessive Exercise	Obsessive routines exceeding healthy limits despite injury or life interference.	no rest day, david goggins mentality, train until failure, push your limit
	Predebting	Treating exercise as punishment or permission to eat, creating guilt cycles.	exercise so I can eat, workout so I can eat, earn your food, train so I can eat
	Maladaptive Coping	Using exercise as the sole coping mechanism to avoid emotional distress.	gym therapy, workout breakup, workout heartbreak, gym mental health, gym fixes everything
	Exercise-Induced Functional Impairment	Prioritizing exercise over essential duties, harming daily functioning.	gym over everything, gym or nothing, gym over friends, skip school for gym
	Toxic Motivation	Demeaning communication that pressures unrealistic fitness standards via shaming or slurs.	gym masculinity, aggressive gym motivation, they don't know me son, get your ass to the gym
Supplements	Anabolic Steroids	Normalization or endorsement of anabolic-androgenic steroid use with downplayed risks.	tren, anabolic stack, anabolic gear, steroids
	Legal APEDs	Overuse of legal supplements (e.g., creatine, protein, preworkout) beyond recommended doses.	protein powder, whey, creatine, pre workout
	Hormone Therapy	Downplaying risks and spreading misinformation about testosterone replacement therapy.	TRT for gains, testosterone, hormoneboost
Relationship to Masculinity		Links muscle-building and exercise performance to male identity, sexuality, and self-worth.	gym masculinity, be a man gym, embrace masculinity, alpha male gym
Irrelevant		General fitness or lifestyle content unrelated to pro–bigorexia harm.	fyp, tiktok, foryoupage, viral, funny, duet, trending, love, meme, followme, repost, new, awesome, music, cute, video, foryou, fun, diy, ootd, family, lifehack, photography, usa, college, travel, christmas, sport, party, popular, clip, movie, star, moment, tiktokviral, tiktokfamous, tiktokmusic, likeforfollow, recipe, quotes, TikTokChallenge, memories

Table 12: Primary and secondary categories of the Harm Taxonomy for Pro-Bigorexia Content on TikTok, with definitions and search keywords. revise the categories definition

Video ID	Subcategory	Video Description	Embedded Text	Audio Transcript
7231300164307143979	Relationship to Body	my lil frame	me being a 5'8 male but with lean body	and my little frame and my sweet little girl voice. It exudes something in people that...
7201534759347014958	Relationship to Masculinity	#menshealth #menshealthtips #cynic	hermes.the.cynic	The reason that there is such a void of male role models for young men is because we literally do not have adult males in our population anymore. What do I mean by that? There's no adult males. Like, there's grown men everywhere, but they're not actually physiologically adult human males...
7275865970306977067	Supplement Abuse	this is totally healthy right? #recipes #creatine	@days.of.j	So I told a co-worker I was just taking my creatine in water, just water, because it was unflavored. And they told me you're supposed to take it with like a meal or like sugar so then you can absorb it better...
7363842693211917611	Relationship to Food	Liver King dinner!	@liverking	Pulled beef with hard shelled eggs Mmm It's crazy. It actually tastes like the delicious pulled beef That's in with the garlic oh and the onion and the lime juice booyah, yeah Sweet potatoes, the purple variety...
7342974339907603718	Relationship to Exercise	I am girl	MAKING THE GYM YOUR WHOLE PERSONALITY no train legs	girls who make going to the gym their entire personality it is just so crazy to me I genuinely do not know how you guys do it I could never be one of those girls anyway someone I know just asked me what day it was and I replied with leg day
7498781444408429867	Irrelevant	Two in #shorts #gymshorts	@johnbshop0	ugly, ugly, ugly aura. Bro, if your shorts, especially if you're hitting leg day, are below your knees, you're genuinely cooked, bro. These, I got them here on the TikTok shop. This is a size medium. I'm five foot 10, 165 pounds, and these things fit so good...

Table 13: Example entries for Type classification showing captions and TikTok metadata. The actual videos can be found in the data attachment in the submission platform.

Video ID	Subcategory	Video Description	Embedded Text	Audio Transcript
7478843939676507423	Muscularity Self-objectification	training hard or hardly training ?	sounds of my workout on push day	Yeah, we need some weight, should be a good one today. Bet you gotta hit me fast, you know old...
7231300164307143979	Leanness Self-objectification	my lil frame	me being a 5'8 male but with lean body	and my little frame and my sweet little girl voice. It exudes something in people that...
7365667609435573509	Muscle Dissatisfaction	#gymtok	When u start working out to look better but now u have body dysmorphia	fucking go let's go I guess
7186715652642622762	Rigid Food Rules	I gained 35lbs in 90 days #lifting #gym #bulking	35lbs in 90 days I've considered ever since 7th grade but change	I gained 35 pounds in 90 days. I've considered myself a runner ever since 7th grade...
7332272427986111790	Unsafe Food	Liver King Dinner to start the week	liverking	Mmm, steak and potatoes, sweet potato fritters, Liver King Chef Lionel, that's how he calls them. So good. We got...
7398320286996663595	Cheat Meals	NO MORE.. #gym #cheatday #gains	us after cheat day turned into cheat month	No more cheese fillets. No more McDonald's. No more chicken wings. No more chicken snobby with some sour. No more...
7198309032384417070	Excessive Exercise	Gym rats! #gym #gymtok #mothersoftiktok #thick #newyearsresolutions #muscles #gains #proteinshake #lawenforcement	It's time for us normal gym peeps to get our routines back!!!	Listen, I got my ass the gym today and I didn't want to go. So did you go today? Well...

Table 14: Example entries for Subtype classification showing captions and TikTok metadata (Part 1 of 2). The actual videos can be found in the data attachment in the submission platform.

Video ID	Subcategory	Video Description	Embedded Text	Audio Transcript
7372218805738589445	Predebting Exercise	You CANNOT our train a bad diet #diet #fitness #exercise	I CAN EAT WHATEVER I WANT AS LONG AS I EXERCISE YUP YOU HEARD ME WHILE EXERCISE IS IMPORTANT...	I can eat whatever I want as long as I exercise. You said what? Yup, you heard me. While exercise...
7231695341077138730	Maladaptive Coping	Gym Therapy #gymtherapy #gym #doordie #dobetter #military	@shreddedvets	Gym therapy is a term that was coined by Shredded Vets that describes the psychological effects of working out consistently...
7460046778524617991	Exercise-Induced Functional Impairment	Bradman Best - I would skip school just to go to the gym. I was going to do whatever it took.	I would skip school to guru the gym.	lucky enough like when I was in year 9 and year 10 the Knights were giving me a day off...
7478474531875704094	Toxic Motivation	put in the effort #gym #motivation #fitnessmotivation #workout	TD PRESSURE	It's fucking different. Gotta really come kill me. I'm built different. I train different. I work different. I am different...
7313287497650326816	Anabolic Steroids	#samsulek	POV: Steroids gives you acne HAVE FUCKING I MAN LIKE PIMPLE SHIT COLLAB MIGHT PIMPLE UH ASIAN	dude I think I have like a fucking I got like a cyst man you know like where it's a...
7213153335786884394	Legal APEDs	Sustenanceee #fyp #xyzbca #gymtok #creatine	POV: you get caught using creatine	Are there any drugs in this house? If there are, you better find them and give them to me immediately...
7190162324563496234	Hormone Therapy	Day 5 on TRT #trt #lift #weightlifting #gym #workout	Day 5 on TRT	What up guys, it's day five. So on day five, the doctor calls you with your full blood work done...

Table 15: Example entries for Subtype classification showing captions and TikTok metadata (Part 2 of 2). The actual videos can be found in the data attachment in the submission platform.