GLAMI-1M:
A Multilingual Image-Text Fashion Dataset

The largest multilingual image-text classification dataset and benchmark

Václav
Košař

Antonín
Hoskovec

Milan
Šulc

Radek
Bartyzal

GLAMI ROSSUM FJFI BMVC 2022

ANKA KEMER Kadın Heybe Çantalı Kemer 16x14 cm
(Turkey)
‘womens-belts‘

Pánská kotníková obuv Mustang 4107-605-820 modrá
(Czechia)
‘mens-boots‘

Ženski kopalni plašč DKaren Basic

(Slovenia)
‘womens-bathrobes‘

Pilgrim Auskarai ’THANKFUL’ sidabrinė
(Lithuania)
‘womens-earrings‘

Multi-Modality and Linguality

Multimodal models:

- can be highly competitive on multiple tasks (CoCa ImageNet SOTA)
- outperform single modality models on (CMA-CLIP on FashionGen)

Multilingual models:

- popular in research and industry (mBERT, XLM-R, mT5)
- multilingual pretraining helps in low resource languages (Conneau 2020)
- machine translation cannot replace human produced text

Large public multilingual multimodal datasets are highly relevant to replicable research.

Pilgrim Auskarai
’THANKFUL’
sidabrinė
(Lithuania)
‘womens-earrings‘

GLAMI-1M Overview

- 1.1M fashion items with 968k unique images and 1.01M unique texts
- 13 languages (cz, sk, gr, hu, si, lt, lv, hr, bg, ee, tr, ro, es)
- 191 fine categories (15 shoes types)
- High-quality annotations
- Difficult e-commerce industry problem

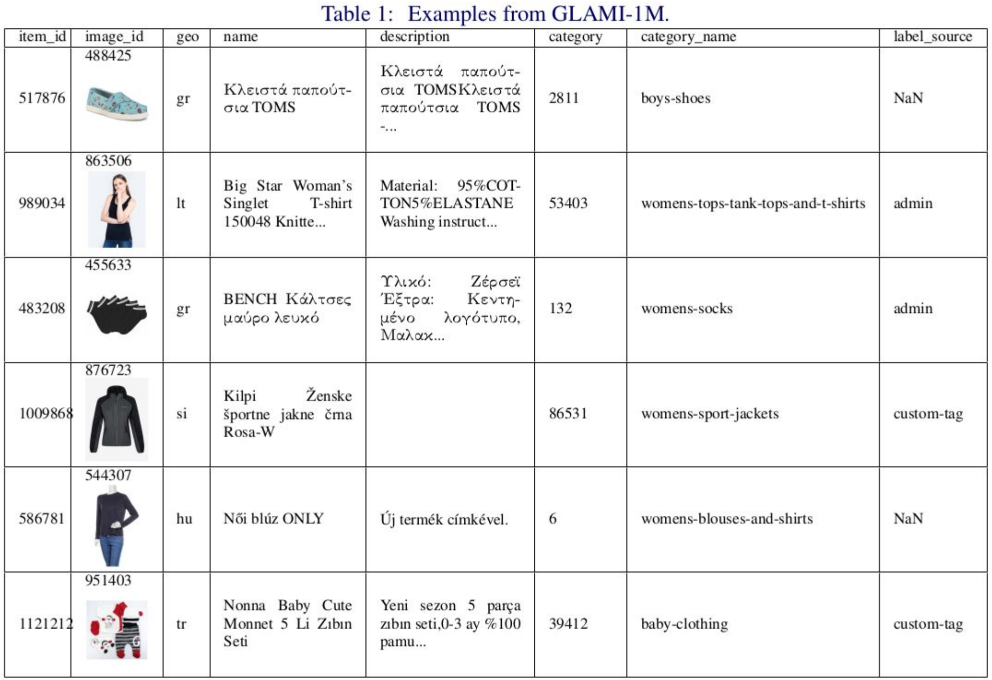
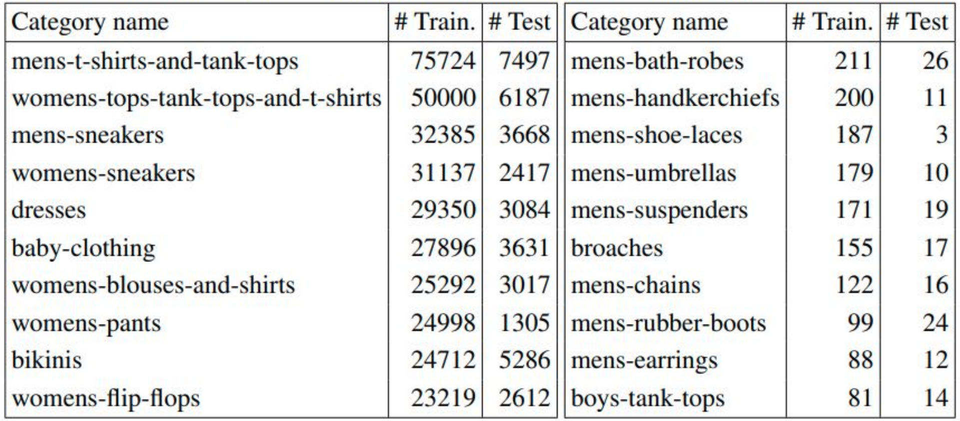


Table 1: Examples from GLAMI-1M.

item_id	image_id	geo	name	description	category	category_name	label_source
517876	488425	gr	Κλειστά παπούτσια TOMS	Κλειστά παπούτσια TOMSΚλειστά παπούτσια TOMS -...	2811	boys-shoes	NaN
989034	863506	lt	Big Star Woman's Singlet T-shirt 150048 Knitte...	Material: 95%COTTON5%ELASTANE Washing instruct...	53403	womens-tops-tank-tops-and-t-shirts	admin
483208	455633	gr	BENCH Κάλτσες μαύρο λευκό	Υλικό: Ζέρσεϊ Έξτρα: Κεντημένο λογότυπο, Μαλακ...	132	womens-socks	admin
1009868	876723	si	Kilpi Ženske športne jakne črna Rosa-W		86531	womens-sport-jackets	custom-tag
586781	544307	hu	Női blúz ONLY	Új termék címkével.	6	womens-blouses-and-shirts	NaN
1121212	951403	tr	Nonna Baby Cute Monnet 5 Li Zıbın Seti	Yeni sezon 5 parça zıbın seti,0-3 ay %100 pamu...	39412	baby-clothing	custom-tag

GLAMI-1M Category Overview

Table 5: The 10 most and 10 least represented from the 191 total training set categories.



Category name	# Train.	# Test	Category name	# Train.	# Test
mens-t-shirts-and-tank-tops	75724	7497	mens-bath-robes	211	26
womens-tops-tank-tops-and-t-shirts	50000	6187	mens-handkerchiefs	200	11
mens-sneakers	32385	3668	mens-shoe-laces	187	3
womens-sneakers	31137	2417	mens-umbrellas	179	10
dresses	29350	3084	mens-suspenders	171	19
baby-clothing	27896	3631	broaches	155	17
womens-blouses-and-shirts	25292	3017	mens-chains	122	16
womens-pants	24998	1305	mens-rubber-boots	99	24
bikinis	24712	5286	mens-earrings	88	12
womens-flip-flops	23219	2612	boys-tank-tops	81	14

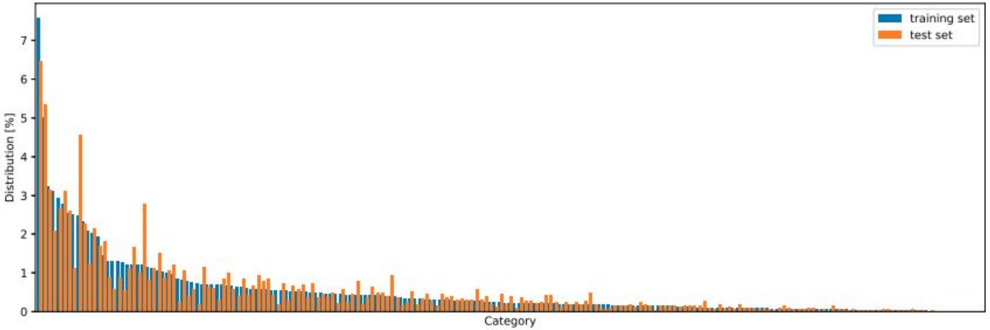


Figure 2: Distribution of samples per category. The distribution is mostly exponential, but steeper along the edges, so we regard this as a long tailed distribution.

GLAMI-1M Classification Dataset Comparison

- The largest multilingual image-text classification dataset.
- The second largest image-text classification dataset - second to Recipe1M+.
- Contrary to Recipe1M+, GLAMI-1M has 75% of the training set, and 100% of the test set human labelled.

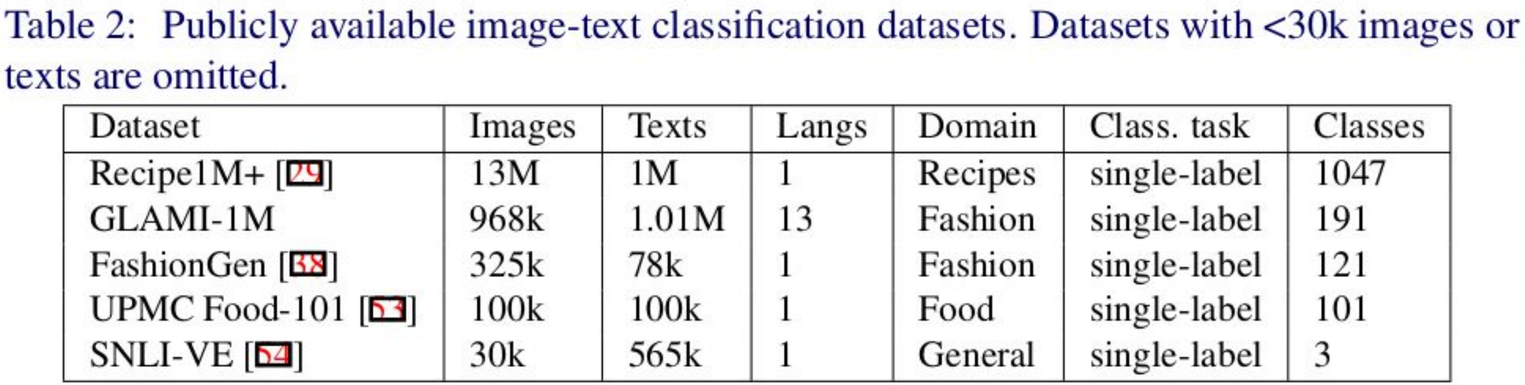


Table 2: Publicly available image-text classification datasets. Datasets with <30k images or texts are omitted.

Dataset	Images	Texts	Langs	Domain	Class. task	Classes
Recipe1M+ [29]	13M	1M	1	Recipes	single-label	1047
GLAMI-1M	968k	1.01M	13	Fashion	single-label	191
FashionGen [38]	325k	78k	1	Fashion	single-label	121
UPMC Food-101 [53]	100k	100k	1	Food	single-label	101
SNLI-VE [54]	30k	565k	1	General	single-label	3

GLAMI-1M Fashion Dataset Comparison

- Image-text fashion dataset with the most items (1.1M).
- The finest grained categories (191), e.g.: 15 shoes types
- The most languages (13).

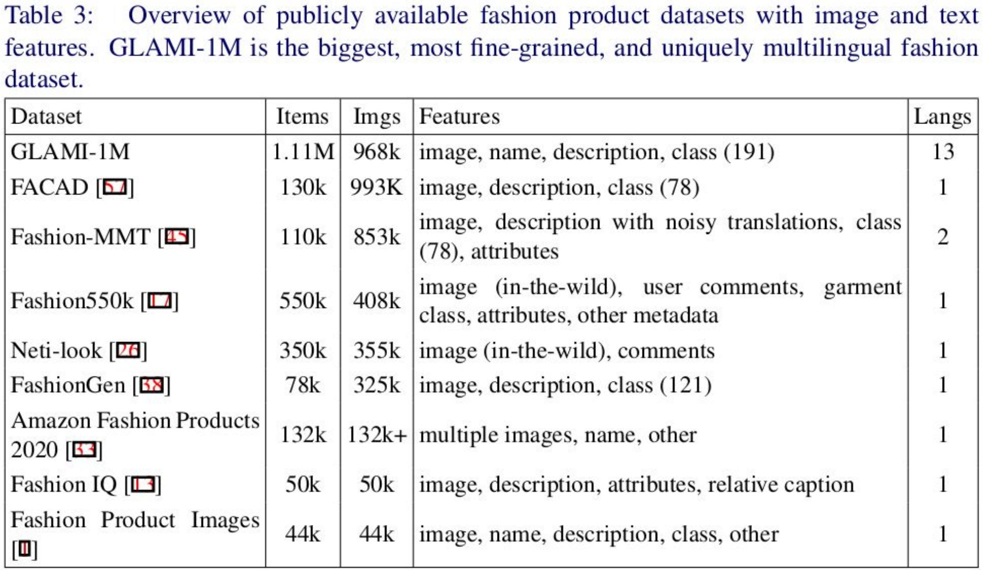


Table 3: Overview of publicly available fashion product datasets with image and text features. GLAMI-1M is the biggest, most fine-grained, and uniquely multilingual fashion dataset.

Dataset	Items	Imgs	Features	Langs
GLAMI-1M	1.11M	968k	image, name, description, class (191)	13
FACAD [57]	130k	993K	image, description, class (78)	1
Fashion-MMT [45]	110k	853k	image, description with noisy translations, class (78), attributes	2
Fashion550k [17]	550k	408k	image (in-the-wild), user comments, garment class, attributes, other metadata	1
Neti-look [26]	350k	355k	image (in-the-wild), comments	1
FashionGen [58]	78k	325k	image, description, class (121)	1
Amazon Fashion Products 2020 [33]	132k	132k+	multiple images, name, other	1
Fashion IQ [13]	50k	50k	image, description, attributes, relative caption	1
Fashion Product Images [1]	44k	44k	image, name, description, class, other	1

GLAMI-1M Web Dataset Comparison

Smaller than web-scale multilingual image-text retrieval datasets, but GLAMI-1M has human classification labels.

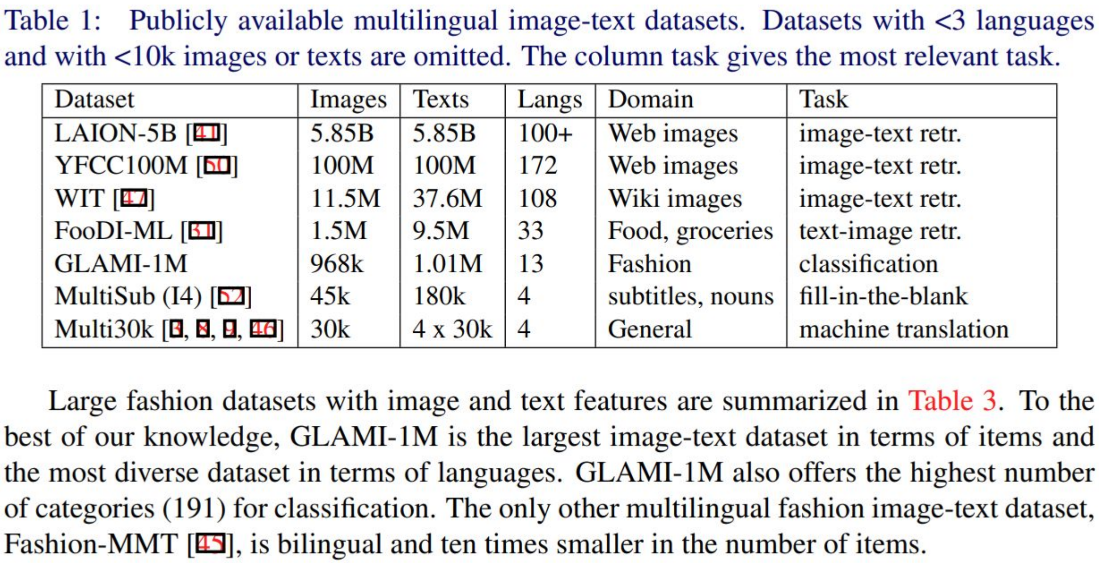


Table 1: Publicly available multilingual image-text datasets. Datasets with <3 languages and with <10k images or texts are omitted. The column task gives the most relevant task.

Dataset	Images	Texts	Langs	Domain	Task
LAION-5B [41]	5.85B	5.85B	100+	Web images	image-text retr.
YFCC100M [50]	100M	100M	172	Web images	image-text retr.
WIT [47]	11.5M	37.6M	108	Wiki images	image-text retr.
FooDI-ML [31]	1.5M	9.5M	33	Food, groceries	text-image retr.
GLAMI-1M	968k	1.01M	13	Fashion	classification
MultiSub (I4) [52]	45k	180k	4	subtitles, nouns	fill-in-the-blank
Multi30k [3, 8, 9, 46]	30k	4 x 30k	4	General	machine translation

Large fashion datasets with image and text features are summarized in Table 3. To the best of our knowledge, GLAMI-1M is the largest image-text dataset in terms of items and the most diverse dataset in terms of languages. GLAMI-1M also offers the highest number of categories (191) for classification. The only other multilingual fashion image-text dataset, Fashion-MMT [45], is bilingual and ten times smaller in the number of items.

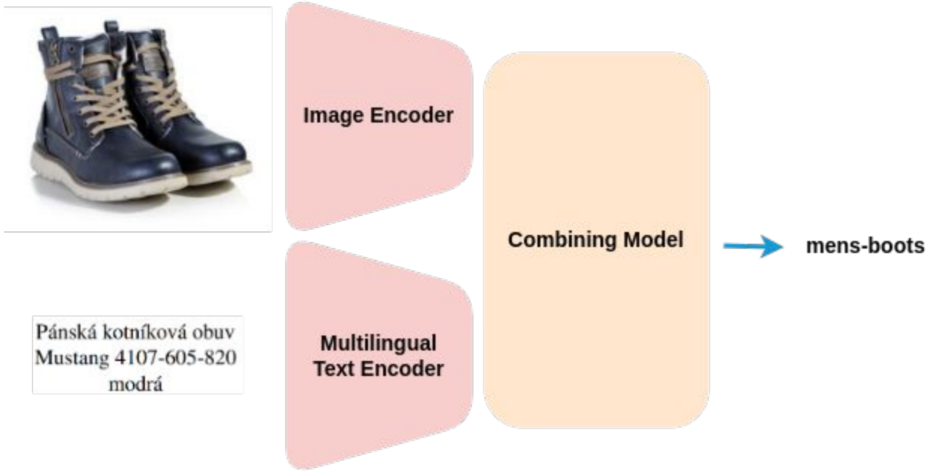
GLAMI-1M
Baseline Models

Publish baseline models for tasks:

- image-text classification: EmbraceNet and zero-shot CLIP
- text-to-image generation: Imagen-like diffusion model

Setup benchmark for the image-text classification on PapersWithCode

Experiments with machine translation.


Pánská kotníková obuv Mustang 4107-605-820 modrá
Image Encoder
Multilingual Text Encoder
Combining Model
mens-boots

GLAMI-1M EmbraceNet Classification Baseline

- Baseline model EmbraceNet consumes mT5 and ResNeXt-50 embeddings inputs and predicts a class
- Non-human training set labels help very little in this setup
- Challenging benchmark since performance only 69.7%

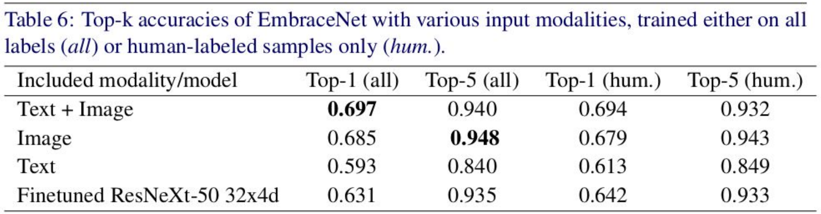
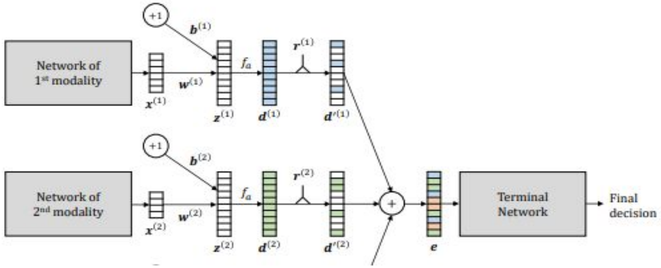


Table 6: Top-k accuracies of EmbraceNet with various input modalities, trained either on all labels (*all*) or human-labeled samples only (*hum.*).

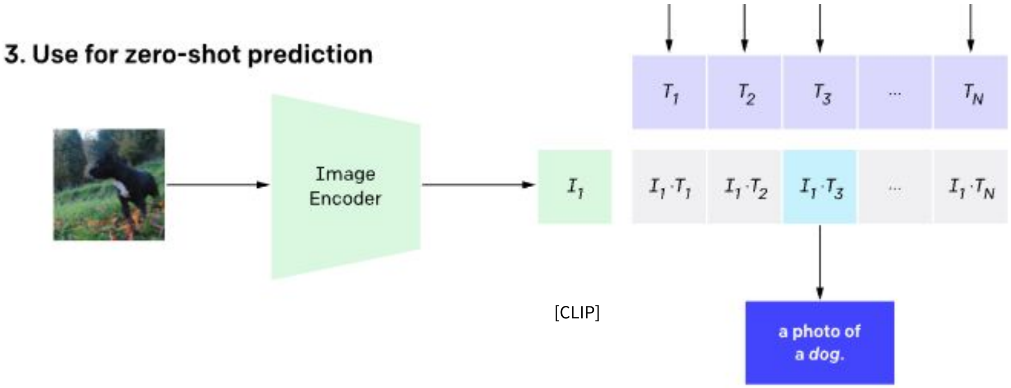
Included modality/model	Top-1 (all)	Top-5 (all)	Top-1 (hum.)	Top-5 (hum.)
Text + Image	**0.697**	0.940	0.694	0.932
Image	0.685	**0.948**	0.679	0.943
Text	0.593	0.840	0.613	0.849
Finetuned ResNeXt-50 32x4d	0.631	0.935	0.642	0.933



[EmbraceNet]

[EmbraceNet] Choi, Jun-Ho and Jong-Seok Lee. “EmbraceNet for activity: a deep multimodal fusion architecture ...” *2019 ACM*

GLAMI-1M CLIP Classification Baseline



[CLIP]

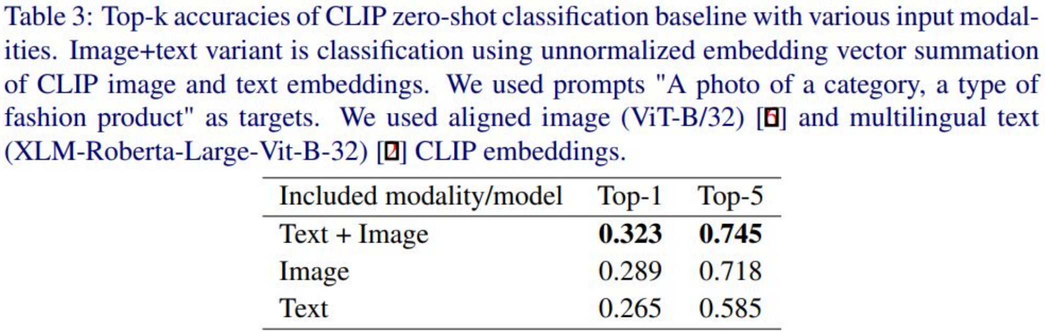


Table 3: Top-k accuracies of CLIP zero-shot classification baseline with various input modalities. Image+text variant is classification using unnormalized embedding vector summation of CLIP image and text embeddings. We used prompts "A photo of a category, a type of fashion product" as targets. We used aligned image (ViT-B/32) [6] and multilingual text (XLM-Roberta-Large-Vit-B-32) [2] CLIP embeddings.

Included modality/model	Top-1	Top-5
Text + Image	**0.323**	**0.745**
Image	0.289	0.718
Text	0.265	0.585

[CLIP] Radford, Alec et al. "Learning Transferable Visual Models From Natural Language Supervision." *ICML* (2021).

GLAMI-1M Text-to-Image Generation Baseline

- Single GPU model
- Multilingual text conditioned cascading diffusion model to 128x128 pixel image
- mT5 text embeddings and two UNet models in a sequence

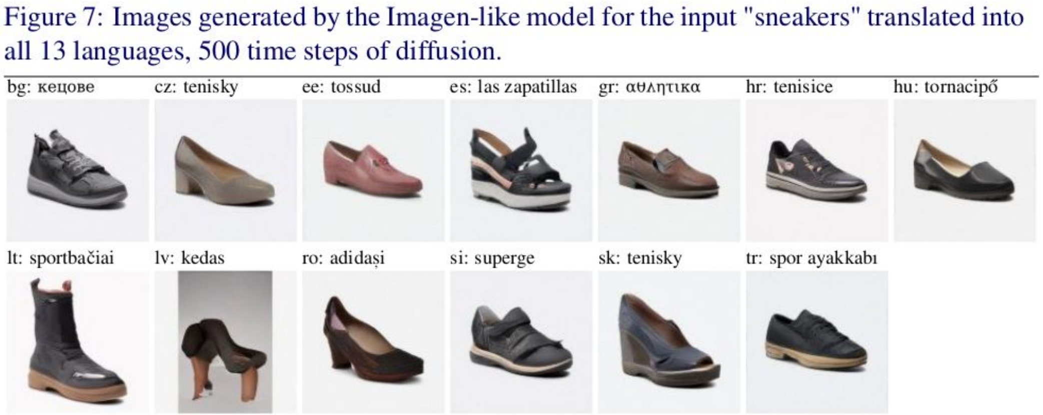


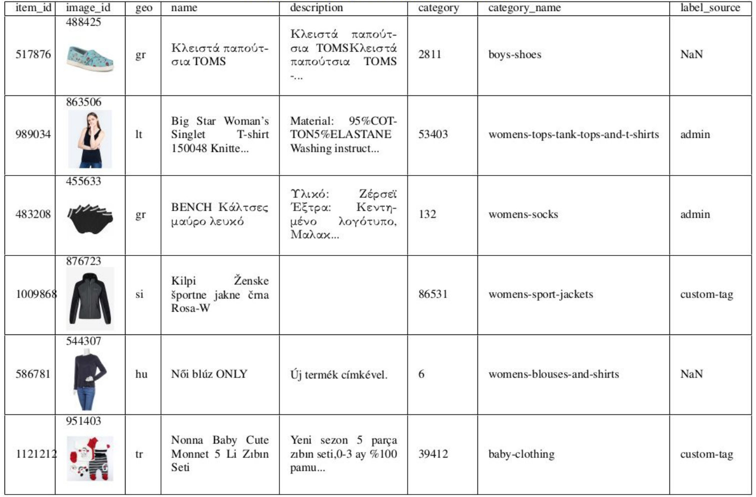
Figure 7: Images generated by the Imagen-like model for the input "sneakers" translated into all 13 languages, 500 time steps of diffusion.

GLAMI-1M Conclusion

The largest multilingual image-text classification dataset and benchmark

- Accelerate research:
- Multilingual alternative to Recipe1M+
- Larger alternative to FashionGen
- Challenging image-text classification benchmark
- Multilingual text-to-image dataset
- Future work: long-tail learning, adaptation to prior shift, learning from a combination of trusted (human) and noisy (rule-based) annotations.

Table 1: Examples from GLAMI-1M.



item_id	image_id	geo	name	description	category	category_name	label_source
517876	488425	gr	Κλειστά παπούτσια TOMS	Κλειστά παπούτσια TOMSΚλειστά παπούτσια TOMS -...	2811	boys-shoes	NaN
989034	863506	lt	Big Star Woman's Singlet T-shirt 150048 Knitte...	Material: 95%COTTON5%ELASTANE Washing instruct...	53403	womens-tops-tank-tops-and-t-shirts	admin
483208	455633	gr	BENCH Κάλτσες μαύρο λευκό	Υλικό: Ζέρσεϊ Έξτρα: Κεντημένο λογότυπο, Μαλακ...	132	womens-socks	admin
1009868	876723	si	Kilpi Ženske športne jakne črna Rosa-W		86531	womens-sport-jackets	custom-tag
586781	544307	hu	Női blúz ONLY	Új termék címkével.	6	womens-blouses-and-shirts	NaN
1121212	951403	tr	Nonna Baby Cute Monnet 5 Li Zıbın Seti	Yeni sezon 5 parça zıbın seti,0-3 ay %100 pamu...	39412	baby-clothing	custom-tag

Thank you!
Any Questions?

- To download the paper or dataset or contact us at: https://github.com/glami/glami-1m
- Download and start using the dataset in your research today.
- Beat our baseline with your own model!

GLAMI

Thank you