RusCode: Russian Cultural Code Benchmark for Text-to-Image Generation

Viacheslav Vasilev[1,2], **Julia Agafonova**[1,3], **Nikolai Gerasimenko**[1], **Alexander Kapitanov**[4], **Polina Mikhailova**[4], **Evelina Mironova**[1], **Denis Dimitrov**[1,5]

[1]Sber AI, [2]MIPT, [3]ITMO University, [4]SberDevices, [5]AIRI

Correspondence: vasilev.va@phystech.edu

Abstract

Text-to-image generation models have gained popularity among users around the world. However, many of these models exhibit a strong bias toward English-speaking cultures, ignoring or misrepresenting the unique characteristics of other language groups, countries, and nationalities. The lack of cultural awareness can reduce the generation quality and lead to undesirable consequences such as unintentional insult, and the spread of prejudice. In contrast to the field of natural language processing, cultural awareness in computer vision has not been explored as extensively. In this paper, we strive to reduce this gap. We propose a RusCode benchmark for evaluating the quality of text-to-image generation containing elements of the Russian cultural code. To do this, we form a list of 19 categories that best represent the features of Russian visual culture. Our final dataset consists of 1250 text prompts in Russian and their translations into English. The prompts cover a wide range of topics, including complex concepts from art, popular culture, folk traditions, famous people's names, natural objects, scientific achievements, etc. We present the results of a human evaluation of the side-by-side comparison of Russian visual concepts representations using popular generative models.

1 Introduction

In recent years, text-to-image (T2I) generation models have achieved a high level of photorealism and comprehension of complex textual prompts (Betker et al., 2023; Esser et al., 2024; Kastryulin et al., 2024; Arkhipkin et al., 2024; Vladimir et al., 2024). This has significantly expanded the potential for using them in various applications, such as advertising, design, education, and art. As these models work with both visual and textual concepts, their operation is closely related to various aspects of human culture. The increasing popularity of generative systems available to users worldwide means that models need to understand text prompts containing specific elements from various cultures. However, as a general rule, these models are trained using large open datasets or data collected from the Internet. Due to the widespread influence of English-speaking popular culture, there is a lack of cultural understanding of other geographical, national, and social groups among generation models (Basu et al., 2023; Qadri et al., 2023; Mim et al., 2024). These restrictions will undoubtedly lead to incorrect generation results for specific cultural concepts, a loss of user interest, and a limited applicability of the model for real-world tasks. In the worst-case scenario, this could lead to undesirable social outcomes, such as unintentional insults (Ghosh et al., 2024), inciting hostility, spreading misinformation, and perpetuating stereotypes and social biases (Naik and Nushi, 2023; Cho et al., 2023; Luccioni et al., 2024). This is one of the main reasons for many concerns regarding the use of generative Artificial Intelligence (AI) in general (Weidinger et al., 2023; Bird et al., 2023).

As human culture is linked to language, similar challenges have been considered previously in natural language processing (NLP) (Hershcovich et al., 2022; Yu et al., 2023; Cao et al., 2024a). However, the cultural awareness issue in visual generation tasks remains largely unexplored. We understand cultural awareness in the generation model as knowledge of the cultural code. We mean the cultural code as a diverse set of concepts that members of a particular social group or nationality regularly encounter. These concepts are often an integral part of a person's cultural background and are widely accepted for communication within specific communities (Corner, 1980). At the same time, these concepts may be unfamiliar or even incomprehensible to other people, as they can contain complex, metaphorical, and fantastical elements.

In this paper, we present the RusCode benchmark dataset for evaluating the quality of image

Findings of the Association for Computational Linguistics: NAACL 2025, pages 7656–7672
April 29 - May 4, 2025 ©2025 Association for Computational Linguistics

generation based on textual descriptions that include concepts from the Russian culture. We conduct cultural analysis with the participation of experts from various fields of the humanities, such as history, literature, sociology, psychology, and philology. Based on these diverse perspectives, we create a list of 19 categories that cover various aspects of Russian culture. We aim to develop a system of concepts that will be easily understood by most native Russian speakers. As a result, we construct a dataset consisting of 1250 complex textual descriptions in Russian and English, which reflect the contextual use of many concepts from traditional and modern Russian culture. When creating the prompts, we took into account the opinions of 13 people from various backgrounds, professions, and age groups. These descriptions include historical, artistic, folkloric, natural, technical and other elements. We also associate a real reference image of a particular entity with each prompt. These images can be used to evaluate the generation quality for the elements of Russian culture with the participation of people who are unfamiliar with Russian culture in detail. This allows one to use our dataset to evaluate multicultural image generation models. We use the collected prompts to generate images using popular models such as Stable Diffusion 3 (Esser et al., 2024), DALL-E 3 (Betker et al., 2023), Kandinsky 3.1 (Arkhipkin et al., 2024; Vladimir et al., 2024), and YandexART 2 (Kastryulin et al., 2024). The results of a human evaluation of the side-by-side comparison of these models provide insight into the current state of multicultural understanding in the modern state of T2I generation.

Thus, the contribution of our work is as follows:

- We analyze the concept of the Russian visual cultural code and create a list of categories that represent the basic cultural background of a native speaker within the context of Russian culture;
- We present a RusCode benchmark dataset of textual descriptions of Russian cultural concepts that can be used to assess the cultural awareness of text-to-image models[1];
- We report on the human evaluation results of the side-by-side comparison of 4 popular text-to-image generation models using collected prompts.

2 Related Works

2.1 Cultural Awareness of Generation Models

We define that a model has a high level of *cultural awareness* if it can generate semantically correct results for text prompts that contain specific concepts related to a particular culture. Multicultural awareness involves understanding of linguistic and semantic features (Wibowo et al., 2023), as well as correctly semantic matching of concepts from different cultures (Cao et al., 2024b). Earlier, a tendency towards Western culture in generative models has been noted (Bhatia et al., 2024; Naik and Nushi, 2023; Berg et al., 2022). As language is a conduit of culture (Ventura et al., 2023), many NLP studies have focused on the issue of cultural awareness. This includes the task of adaptive translation (Peskov et al., 2021), offensive language detection (Zhou et al., 2023; Awal et al., 2024), dialog systems operation (Cao et al., 2024a) and other tasks (Hershcovich et al., 2022). The development of visual-language models (VLM) has led to a transfer of cultural awareness issues for the multimodal architectures (Nayak et al., 2024). This problem was considered in the context of the visual question answering (VQA) task (Becattini et al., 2023; Romero et al., 2024), image-text retrieval and grounding (Bhatia et al., 2024). With the addition of a new modality, the problem of cultural awareness has become more acute. For example, there are different levels of understanding of concepts from regional cultures among modern VLMs (Nayak et al., 2024). In text-to-image generation, quality metrics have long focused on the aesthetics and photorealism of generated results, while ignoring cultural awareness (Kannen et al., 2024) Several studies have identified significant gaps in the level of multicultural awareness for the most popular T2I models. As far as we know, our work represents the first comprehensive approach to the issue of cultural awareness in relation to Russian culture in the T2I task.

2.2 Multicultural Benchmarks

Benchmarks for evaluating the multicultural and multilingual abilities of generative models primarily emerged in NLP tasks. Due to the fact that most existing language models are designed primarily for English, several studies have focused on assessing the linguistic and grammatical features of other languages (Cahyawijaya et al., 2021; Zhang et al., 2022; Mukherjee et al., 2024; Kim et al.,

[1] Dataset is available here: `https://github.com/ai-forever/RusCode`

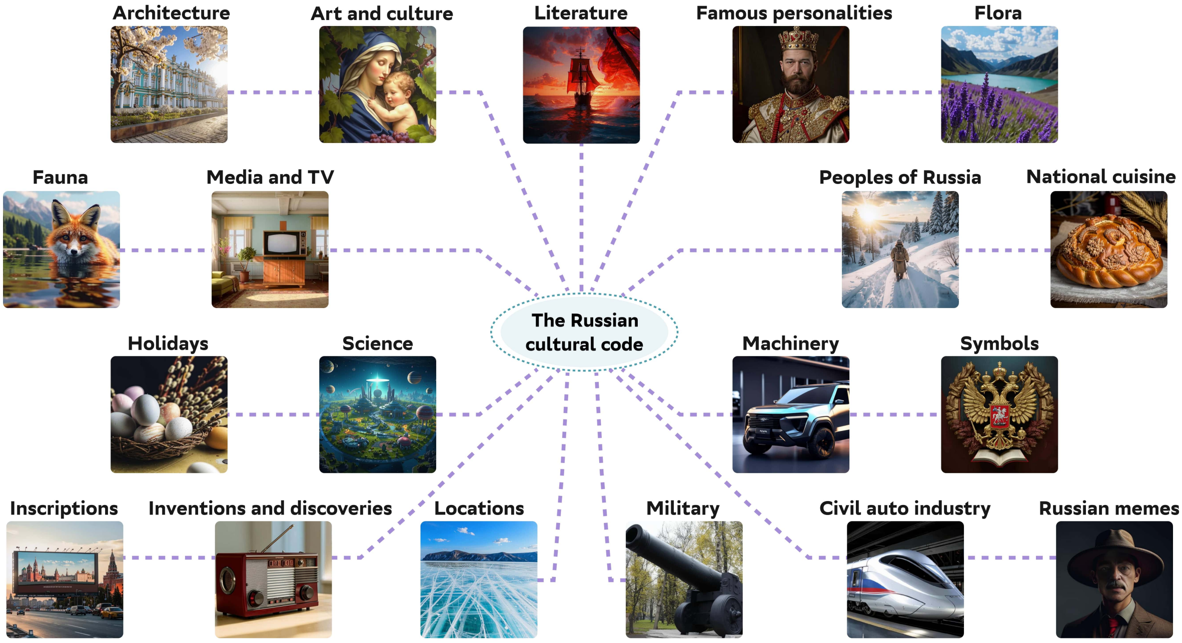


Figure 1: 19 categories of Russian cultural code in our RusCode benchmark dataset. The images are generated by the Kandinsky 3.1 model (Arkhipkin et al., 2024; Vladimir et al., 2024).

2024; Fenogenova et al., 2024; Taktasheva et al., 2024). These benchmarks were designed to expand the range of multilingual knowledge tested in language models, as previously translated English-language datasets have overlooked various cultural and linguistic features (Kim et al., 2024). This is especially important for common languages, which, nevertheless, have limited resources in terms of accessible open information on the Internet (Cahyawijaya et al., 2021; Zhang et al., 2022). The main types of tasks included in such benchmarks were question answering (Kim et al., 2024), natural language generation (Cahyawijaya et al., 2021), multilingual dialog generation (Zhang et al., 2022), text style transfer (Mukherjee et al., 2024), and many other tasks (Fenogenova et al., 2024).

The next significant step forward was the development of multimodal benchmarks for evaluating multilingual VLMs (Liu et al., 2022; Bugliarello et al., 2022; Nayak et al., 2024; Romero et al., 2024; Inoue et al., 2024). The range of benchmark tasks here primarily includes visual question answering (VQA) (Bugliarello et al., 2022; Nayak et al., 2024; Romero et al., 2024; Inoue et al., 2024), as well as cross-modal retrieval, grounded reasoning, and grounded entailment tasks (Bugliarello et al., 2022). Among the findings regarding the results of applying these benchmarks, it has been noted that the quality of modern VLM models varies depending on geographic and cultural categories (Nayak et al., 2024). In addition, efforts have been made to combine text collected from the Internet with text generated by a pre-trained image captioning model.

Due to the fact that the T2I task primarily requires an assessment of visual cultural characteristics, not much work has been done in this area. Currently, existing benchmarks are limited in terms of the number of languages and cultural categories they cover (Kannen et al., 2024).In addition, they do not support the Russian language, despite its relatively high level of usage on the Internet[2]. In this work, we are, to the best of our knowledge, the first to conduct a comprehensive cultural analysis in order to create a benchmark dataset for assessing the quality of image generation incorporating elements of Russian culture.

2.3 Ethics and Social Biases in Generative AI

Insufficient cultural awareness of image generation models can lead to the spread of social biases, misinformation, and offensive content (Naik and Nushi, 2023; Cho et al., 2023; Luccioni et al., 2024). A number of studies have focused on reducing the biases in generative models that are based on factors such as race, skin color, gender, geogra-

[2] https://w3techs.com/technologies/overview/content_language

Table 1: The list of categories and subcategories in the RusCode benchmark dataset

Categories	**Subcategories**
Architecture	Orthodox Church; Sights; Major cities
Art and culture	Painting; Music; Theater; Ballet; Opera; Musical; Photography; Cinema; Cartoons; Architecture; Sculpture; Decorative and Applied arts; Design; Circus
Literature	Folklore, fairy tales and legends; Poems; Prose; Fables; Children's literature
Famous personalities	Public figures; Cultural figures; Scientists; Entrepreneurs; Military; Cosmonauts; Russian writers; Musicians; Actors; Bloggers; Politicians; Athletes
Flora	Coniferous plants; Deciduous plants; Trees and shrubs; Flowers and herbs; Tundra vegetation; Steppe vegetation; Swamp vegetation; Desert vegetation; Fungi; Lower plants; Spore plants; Fruit plants; Berries; Root crops
Fauna	Mammals; Fish; Birds; Reptiles; Cold-blooded; Artiodactyls; Ungulates; Carnivores; Herbivores; Amphibians; Wild animals; Domesticated animals; Small animals; Large animals
Media and TV	Animation; Documentaries; TV Series; Talk Shows; Reality Shows; Feature films; Social networks; Advertising
Peoples of Russia	Nationalities; Clothing; Traditions; Religion; Crafts
National cuisine	First courses; Second courses; Hot appetizers; Cold appetizers; Desserts; Meat dishes; Fish dishes; Milk and dairy products; Bread and bakery products; Cereals; Vegetables; Fruits; Soft drinks; Alcoholic drinks
Holidays	Religious holidays; Civil holidays; Political holidays; Family holidays; Professional holidays; National holidays; International holidays
Science	Natural Sciences; Exact Sciences; Social and Humanitarian Sciences; Fundamental Sciences; Applied Sciences
Machinery	Modern machinery; Soviet machinery; Agricultural machinery; Aviation machinery; Shipping machinery; Construction machinery
Symbols	State symbols; National symbols
Inscriptions	Signage and billboards; Logos and symbols
Inventions and discoveries	
Russian memes	
Locations	Natural; Man-made
Civil auto industry	Passenger cars; Trucks; Public transport
Military technics	Tanks, armored personnel carriers, air defense; Airplanes and helicopters; Ships and submarines; The rest; Equipment of the 19th century; Equipment of the 17th-18th century; Equipment of an earlier period

phy, and social status (Dehouche, 2021; Naik and Nushi, 2023; Yu et al., 2023; Birhane et al., 2024; Clemmer et al., 2024). Cultural stereotypes were also seen as undesirable in favor of greater globalization (Berg et al., 2022; Struppek et al., 2024). Although we agree that cultural stereotypes can be offensive and need to be eliminated, knowledge about the specific cultural features should be retained by the model. For this reason, in our work we create a benchmark using expert knowledge to test the model's ability to capture real cultural features, while avoiding offensive stereotypes.

3 Cultural Code Analysis

The concept of *cultural code* is a complex idea that draws upon various fields such as history, cultural studies, sociology, philosophy, semiotics, and communication theory. The cultural code of a particular community is formed by symbolic systems such as language, art, as well as traditions, along with norms, values, social practices, and historical background. Popular culture, visual media and other forms of information significantly contribute to shaping the cultural code (Corner, 1980). Generation models need to be trained on a deeper understanding of cultural codes. This would improve the visual quality of their outputs, enhance their interpretative and communicative abilities, and enable the creation of systems that are sensitive to cultural diversity. This in turn would reduce biases and foster a more ethical approach in the content generation.

In this study, we explore the Russian cultural code. Drawing on existing research (Goloubkov, 2013; Billington, 2010; Figes, 2002; Stites, 1992), we highlight language, literature, art, religion, philosophy, folklore, and history as central components of Russian cultural identity. Since language reflects cultural characteristics (Wierzbicka, 2002), it is essential for the model to accurately interpret metaphors, proverbs, and figurative expressions that are common in everyday communication. From a visual standpoint, we also consider elements of contemporary popular culture, such as film and TV, as well as geographical places and natural objects. To ensure that our data selection

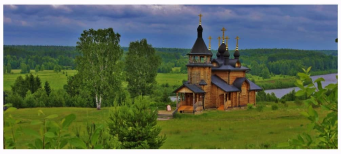
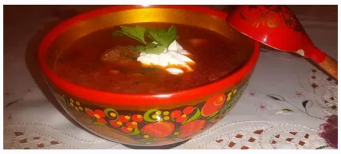
Category	Subcategory	English Prompt	Reference image
Locations	Man-made	Beautiful temple with golden domes at sunset	
Symbols	National symbols	Balalaika lies on a table covered with an embroidered tablecloth	
Peoples of Russia	Crafts	A Russian blacksmith forges a sword for a hero in his forge	
Art and culture	Sculpture	«Motherland» sculpture in Volgograd, against the backdrop of the night sky with firework	
Art and culture	Architecture	Old Russian church in the village	
National cuisine	First courses	Borscht with sour cream in a bowl with Khokhloma painting	

Figure 2: Examples of prompts from RusCode dataset with corresponding reference images

aligns with this cultural framework, we consulted experts from different fields including history, literature, sociology, psychology, and philology. Their collective efforts have resulted identifying of 19 main categories and 125 subcategories, which are crucial for accurately representing the visual dimension of the Russian cultural code. Figure 1 shows these categories with image examples. A complete list of categories and subcategories can be found in Table 1.

4 Dataset

4.1 Prompts Creation

General remarks. After defining a list of main categories and subcategories that represent the Russian cultural code (Section 3), we have created a dataset with prompts that correspond to these subcategories. We assigned ten complex prompts to each of the 125 subcategories in order to ensure a balanced distribution of cultural concepts in the dataset. Each prompt is presented in Russian and has an English translation variant. By a *complex prompt*, we mean a textual description enclosing a certain cultural concept in a specific context of its use. For example, in the prompt "`art photography, aerial view of the Bolshoi Theatre in Moscow, evening, sunset`" two important entities for Russian culture related to each other are mentioned at once – "`The Bolshoi Theater`" and "`Moscow`". At the same time, the dataset contains concepts that are not directly expressed through visual images and require additional creative description. For instance,

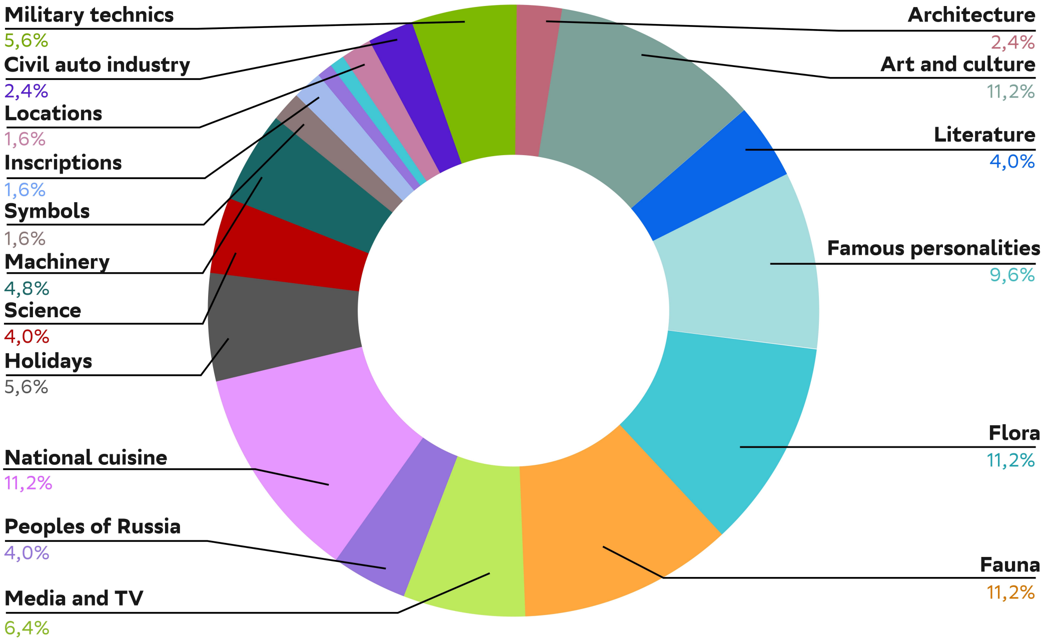


Figure 3: The ratio of the number of collected prompts by each category in the RusCode dataset.

the discovery of an industrial process for producing synthetic rubber is represented in the dataset through the following description: "`Young Soviet chemist Sergei Lebedev made a discovery: a chemical reaction with a substance coming out of a flask, smoke and soot on the scientist's face, the invention of rubber`". More examples of complex prompts are presented in Figure 2.

Prompt-engineers. It was essential for us that the prompts reflect the diverse experiences of people from the Russian culture. We have assembled a team of 13 prompt-engineers, including native speakers and professionals from various backgrounds. It includes a doctor, a manager, a cook, a pharmacist, a translator, an editor, a photojournalist, a psychologist, a car mechanic, a builder, a logistics expert, a linguist, and a copywriter. The age of people ranged from 19 to 46 years old. All prompt-engineers had previous experience in creating textual description datasets for generation models. They were officially employed when they completed their task and were aware of the work's objectives and the possible disclosure of their main areas of activity. Each team member has received instructions and been made aware of the rules for data collection, including ethical considerations and copyright laws.

Prompting. We did not expressly limit the authors in any way at the initial stage of creating prompts. We recommended that they rely on their own experience and imagination. They were also provided with visual examples of how images with the Russian cultural code should look, descriptions of which they should create. Prompt-engineers actively used reference literature, books on painting and history, as well as open resources on the Internet. In order to avoid potential inaccuracies and enrich the dataset with more specific concepts, they did not utilize large language models. The final breakdown of the number of collected prompts by category is presented in Figure 3.

4.2 Prompts Filtering and Post-processing

After the initial collection of prompts for each subcategory, they were selected and filtered by two experienced and professional prompt-engineers, who also contributed to the creation of a list of categories and subcategories (Section 3). During the selection process, they were guided by their experience in creating high-quality and effective prompts, as well as by the idea of what prompts people use when they think about a particular entity and want to generate image with it. Additionally, professional prompt-engineers have corrected and rewritten the descriptions based on the results of popular queries in search engines related to Russian culture. They checked the correctness of the descriptions to

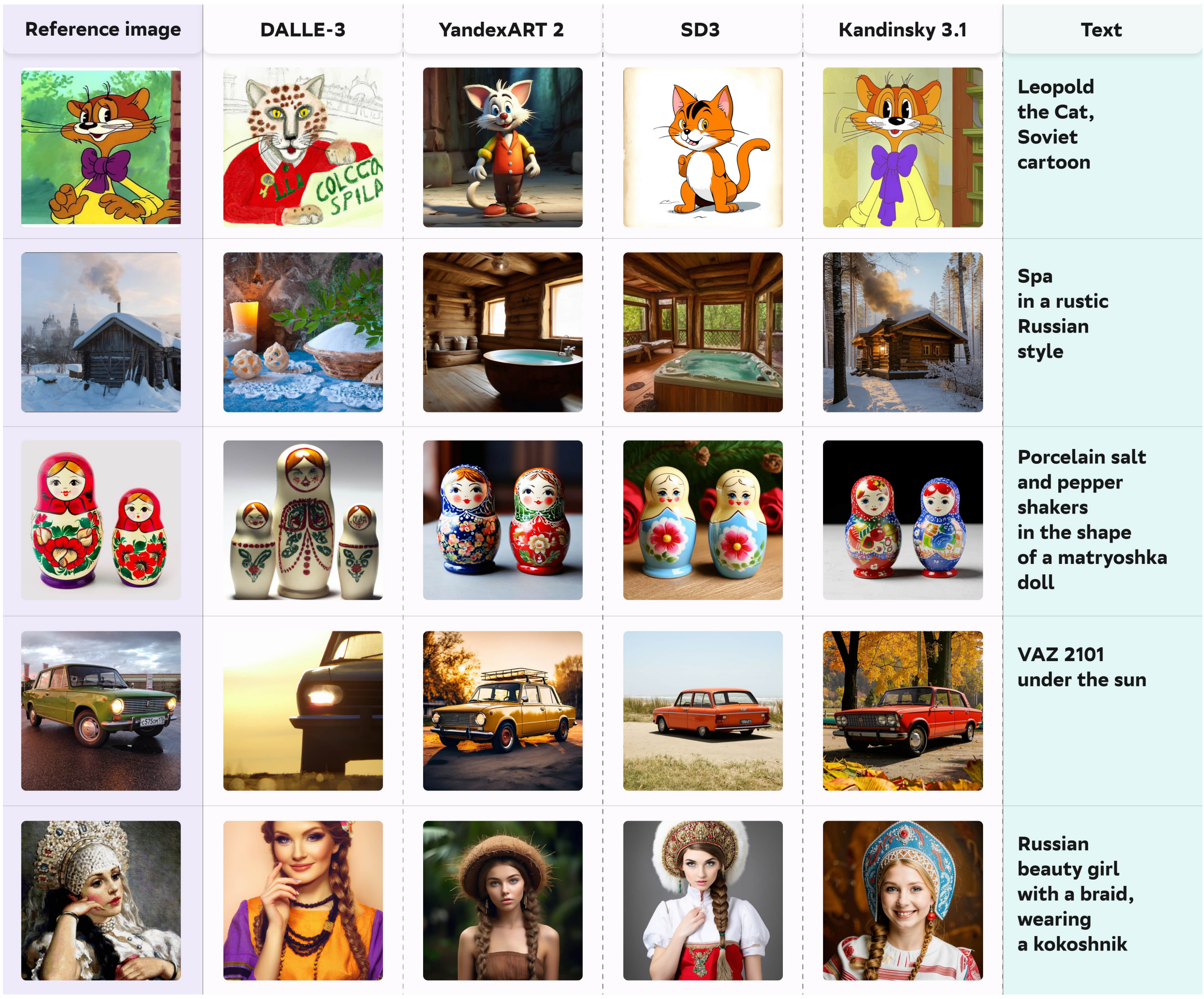

Reference image
DALLE-3
YandexART 2
SD3
Kandinsky 3.1
Text
Leopold the Cat, Soviet cartoon
Spa in a rustic Russian style
Porcelain salt and pepper shakers in the shape of a matryoshka doll
VAZ 2101 under the sun
Russian beauty girl with a braid, wearing a kokoshnik

Figure 4: Comparison of Russian cultural code generations for popular text-to-image models. Reference is an example of a real image with a specific cultural concept from RusCode dataset.

ensure they matched reality and referenced literature, and added a plot to the prompts, enhancing its creativity, variety and detail. Special attention was given to prompts that describe complex and less popular topics, on which there is limited information in open sources or no visual content available. We also used the statistics on popular prompts for T2I generation models. Thus, 1250 prompts were selected from the 2500 initially created.

4.3 Reference Image Collection

We also include high-quality reference images corresponding to each prompt in the dataset. These images are taken from open sources on the Internet. This is necessary to expand the possibilities of using our benchmark to visually assess the correctness of displayed cultural concepts with the help of people unfamiliar with Russian visual culture. By comparing the generated object or entity with the reference image, one can see how correctly the T2I model has performed in generating it. The images were selected after creating a set of prompts. Therefore, choosing images that matched the text descriptions as closely as possible was necessary. The reference images were selected by the same team of people who created the prompts (Section 4.1). Special attention was paid to the aesthetics, photorealism, and overall visual quality of the images, including factors such as contrast, relative positioning of objects, brightness, color saturation, the naturalness of color reproduction, and sharpness. Examples of English-language prompts from different categories and subcategories, along with corresponding reference images, are presented in Figure 2.

5 Evaluation

Qualitative comparison. We used the RusCode dataset to generate images using four popular T2I models, such as Stable Diffusion 3 (Esser et al.,

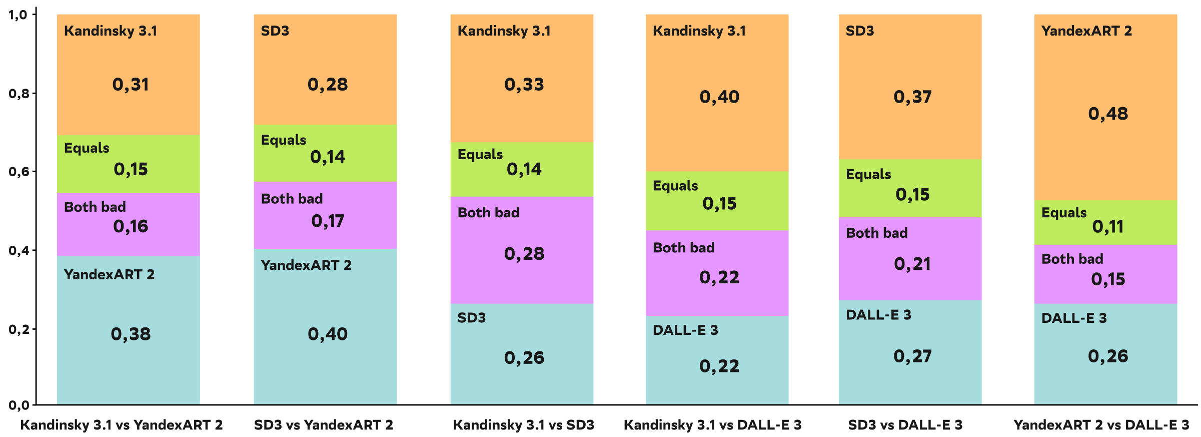


Figure 5: Human evaluation results of a side-by-side comparison between T2I model generations using text prompts from the RusCode dataset.

2024), DALL-E 3 (Betker et al., 2023), Kandinsky 3.1 (Arkhipkin et al., 2024; Vladimir et al., 2024), and YandexART 2 (Kastryulin et al., 2024). Figure 4 shows several examples. As can be seen from the comparison with reference images, although all models have a fairly high level of visual quality, they cope with the generation of Russian cultural entities in different ways. Disadvantages can be expressed both in a lack of complete understanding of a particular entity, as well as in incorrect presentation of details. In some cases, models capture common features and produce examples of generalized concepts composed of individual recognizable elements, but they do not accurately reflect the essence of the reference.

Human evaluation. We compared each of the four models side-by-side with the other three, conducting a human evaluation study. Each person was shown simultaneously the generations of two models without specifying their names. The task was to choose the image that most accurately matches the text description. A team of 48 people who were not involved in the creation of the dataset participated in the evaluation of the generated content. The age range of the participants was between 18 and 54 years. The fields of study and professions of the participants covered information systems, anthropology, programming, law, economics, philosophy, philology, linguistics, regional studies, political science, design, pedagogy, journalism, ecology, finance, sports, management, agriculture, and more. Each person viewed approximately 125 image pairs. The results of a general comparison across all categories are presented in Figure 5. The results of comparing models in individual categories can be found in the Appendix B. As can be seen, the Kandinsky 3.1 and YandexART 2 models significantly outperform the Stable Diffusion 3 and DALL-E 3 models. This indicates a lack of understanding of Russian culture among some of the most popular generative models. The Appendix A contains the results of comparing the Kandinsky 3.1 and Midjourney v6 (Midjourney, 2022) models.

It is important to note that sometimes models blocked generation for some prompts due to excessive self-censorship. For example, the YandexART 2 model blocked the prompts “`opera War and Peace, ball scene`”, “`Monument to the heroes of the Battle of Stalingrad on Mamayev Kurgan`”, “`Ostankino TV Tower at Night`”, “`Mikhail Gorbachev in a hat`”, etc. These and other prompts from our dataset do not contain any real offensive content. As a rule, automatic censorship does not react correctly to the mention of anything related to historical military topics or specific historical figures. When evaluating, we considered such censorship as a “bad” case. The Midjourney v6 model allowed us to use only 974 prompts out of 1250, so we did not include a comparison with it in the main text.

CLIP Score. We used CLIP Score (Radford et al., 2021) to try to automatically assess the cultural awareness of the models on our dataset. The similarity score between the embeddings of the English prompts and the corresponding image embeddings are presented in Table 2. As can be seen, the results for all models are quite high and do not correlate with the human evaluation results. This confirms the inadequacy of using CLIP score for the cultural awareness assessment.

Table 2: The similarity score between the embeddings of text prompts in English from the RusCode dataset and the embeddings of the corresponding generated images.

	CLIP Score ↑
Stable Diffusion 3	26.90
DALL-E 3	27.38
Midjourney v6	27.74
Kandinsky 3.1	26.89
YandexART 2	26.60

6 Discussion

Taxonomy of errors. We identified the features of errors that models encounter in trying to generate something from the Russian cultural code. In the absence of appropriate training examples, the model, even using a sufficiently detailed textual description, will not be able to correctly generate the necessary entity. Nevertheless, for large popular models, we observe a distortion of the entity or a display of an international concept rather than its replacement by an entity from another culture.

Automatic metrics. As far as we know, there are currently no automatic metrics for assessing the cultural awareness of image generation models. The use of automatic metrics such as CLIP-score is not suitable for this task, since the evaluator model itself has a low level of cultural awareness (Table 2). This leads to the need to rely primarily on human evaluation, although we believe that our benchmark can lead to the development of automated tests for this task, for example, based on modern visual-language models, finetuned for cultural specifics.

Causes and mitigation of the cultural awareness gap. We explain the advantage of the Kandinsky 3.1 and YandexArt 2 models by the presence of training data in the domain of Russian culture. The authors of Kandinsky 3.1 write about this explicitly in their technical report (Arkhipkin et al., 2024), while it is not exactly known for YandexArt 2. However, the fact that YandexArt 2 is primarily focused on interacting with Russian users allows us to make such an argumentative assumption. Following Kandinsky 3.1, we think that fine-tuning based on specific culture data will significantly improve the cultural awarness of the model. We also note that retrieval-augmented generation (RAG) methods (Lewis et al., 2020) can be productive in this direction.

7 Conclusion

In this paper, we proposed an open T2I benchmark dataset RusCode, which contains 1250 prompts in Russian and their corresponding translations into English. The dataset will be published under the MIT license. As far as we know, this is the first study in which the Russian cultural code has been examined in such a comprehensive and detailed manner. Despite the complexity of analyzing the national cultural code of any country, we have managed to create a system of categories and subcategories that accurately reflects the basic understanding of average users regarding prompts related to Russian visual concepts. The generation results of popular T2I models proof the existence of a cultural awareness issue, even though, in general, these models have some knowledge of generalized concepts. We strive to expand the use of our dataset. To do this, we attach reference images to the textual prompts, which can be used to mark the correctness of the generated entity. In the future, we aim to attract new experts and significantly increase our dataset, both in terms of the number of categories and the number of prompts in each subcategory. In the future, we also plan to expand this approach for video generation task (Arkhipkin et al., 2023, 2025).

8 Limitations

Incompleteness of categories. The assessment of the visual generation quality of elements of the Russian cultural code, which can be obtained using our dataset, may still not give a complete understanding of the capabilities of the generation model. This is directly related to the complexity and ambiguity of the concept of the Russian cultural code, which we significantly narrow down by providing a specific list of categories. This list could be significantly expanded, but we have focused on the most common concepts among ordinary users. At the same time, we do not exclude the fact that the dataset could contain data that requires more professional knowledge and goes beyond basic erudition or, for example, the school curriculum.

Insufficient representation of individual subcategories. Within each subcategory, we have presented only 10 prompts to balance the data and avoid giving preference to any particular topic. At the same time, the importance of individual subcategories, such as "Prose" within the category of

"Literature", is significant for Russian culture. It can be difficult to determine how the proportion of data in a dataset should reflect the social and cultural significance of a particular topic. Therefore, some relatively important concepts may be overlooked in the dataset, and preference may be given to less significant ones.

Ignoring more subtle differences. In this paper, we use the term "Russian cultural code" to refer to a set of cultural phenomena that are common in Russia and the post-Soviet states. This may lead to some blending of differences between the various peoples and ethnic groups that inhabit the vast territory of Russia. However, our dataset is a significant step towards cultural awareness of these cultures, as the languages and cultural ideas of the people of Russia are closely related. As mentioned earlier (Cahyawijaya et al., 2021), training a model on closely related languages can enhance the quality of results in languages with limited resources. Therefore, we believe our dataset can also capture the cultural traits of the smaller ethnic groups.

Ambiguity and obsolescence. In the humanitarian field, opinions can often lead to disagreements and disputes, and we acknowledge the possibility of mistakes related to cultural nuances. The cultural landscape is constantly evolving, influenced by elements of popular culture, new events, and trends, and we aim to monitor this process by regularly updating our data and expanding our team of experts. We will strive to standardize the process of prompts creation while maintaining the necessary level of freedom for prompt-engineers.

Quality assessment. In this paper, we rely on human evaluation, based on a side-by-side comparison of model generation results for collected prompts. In the future, we plan to expand the list of quality metrics to include realism and visual quality assessment.

Recommendations for use. Although we provide reference images, there may still be an incorrect match between the reference entity and the generated one when testing with people unfamiliar with Russian culture. We also recommend involving experts in relevant fields and using reference literature from trusted sources in such assessments.

9 Ethical Statement

Dataset Content. We have avoided any potentially offensive or discriminatory concepts in our dataset. We do not include any racial prejudices or historical elements that might indicate a biased attitude towards any group in the concept of the cultural code. We strongly oppose nationalism and xenophobia. However, some elements of traditional culture might conflict with modern views of individuals or social groups. It is important to treat this with an understanding of their historical and cultural context.

Personal data. Our dataset contains names and portraits of well-known historical figures of Russian culture. We would like to emphasize that we have not violated their privacy, as all information and images used in our dataset were obtained from open sources.

Usage. Our research aims to promote multiculturalism and diversity in artificial intelligence. We oppose using our data for any illegal purposes, including incitement of hostility, hatred, or the creation of technologies that misinform, create false, or politically biased materials.

Payment for prompt-engineers and evaluators. We properly paid the work of prompt-engineers who participated in the collection of the dataset (Section 4.1), and the participants of the human evaluation study (Section 5). The average salary for each person exceeded the average salary in the city of his residence, according to publicly available government statistics.

Acknowledgments

The authors express their gratitude to Tatiana Nikulina, Anastasia Alyaskina, Denis Kondratiev, Mikhail Kornilin, Julia Filippova, Irina Tabunova, Alexandra Averina, Stefaniya Kozlova, Andrei Filatov as well as Tagme and ABC Elementary Markup Commands.

References

Vladimir Arkhipkin, Andrei Filatov, Viacheslav Vasilev, Anastasia Maltseva, Said Azizov, Igor Pavlov, Julia Agafonova, Andrey Kuznetsov, and Denis Dimitrov. 2024. Kandinsky 3.0 technical report. *Preprint*, arXiv:2312.03511.

Vladimir Arkhipkin, Zein Shaheen, Viacheslav Vasilev, Elizaveta Dakhova, Andrey Kuznetsov, and Denis

Dimitrov. 2023. Fusionframes: Efficient architectural aspects for text-to-video generation pipeline. *Preprint*, arXiv:2311.13073.

Vladimir Arkhipkin, Zein Shaheen, Viacheslav Vasilev, Elizaveta Dakhova, Konstantin Sobolev, Andrey Kuznetsov, and Denis Dimitrov. 2025. Improveyourvideos: Architectural improvements for text-to-video generation pipeline. *IEEE Access*, 13:1986–2003.

Md Rabiul Awal, Roy Ka-Wei Lee, Eshaan Tanwar, Tanmay Garg, and Tanmoy Chakraborty. 2024. Model-agnostic meta-learning for multilingual hate speech detection. *IEEE Transactions on Computational Social Systems*, 11(1):1086–1095.

Abhipsa Basu, R. Venkatesh Babu, and Danish Pruthi. 2023. Inspecting the geographical representativeness of images from text-to-image models. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 5113–5124.

Federico Becattini, Pietro Bongini, Luana Bulla, Alberto Del Bimbo, Ludovica Marinucci, Misael Mongiovì, and Valentina Presutti. 2023. Viscounth: A large-scale multilingual visual question answering dataset for cultural heritage. *ACM Trans. Multimedia Comput. Commun. Appl.*, 19(6).

Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 806–822, Online only. Association for Computational Linguistics.

James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, Wesam Manassra, Prafulla Dhariwa, Casey Chu, Yunxin Jiao, and Aditya Ramesh. 2023. Improving image generation with better captions.

Mehar Bhatia, Sahithya Ravi, Aditya Chinchure, Eunjeong Hwang, and Vered Shwartz. 2024. From local concepts to universals: Evaluating the multicultural understanding of vision-language models. *Preprint*, arXiv:2407.00263.

James Billington. 2010. *The icon and axe: An interpretative history of Russian culture*. Vintage.

Charlotte Bird, Eddie Ungless, and Atoosa Kasirzadeh. 2023. Typology of risks of generative text-to-image models. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 396–410, New York, NY, USA. Association for Computing Machinery.

Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. 2024. The dark side of dataset scaling: Evaluating racial classification in multimodal models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '24, page 1229–1244, New York, NY, USA. Association for Computing Machinery.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. IGLUE: A benchmark for transfer learning across modalities, tasks, and languages. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pages 2370–2392. PMLR.

Samuel Cahyawijaya, Genta Indra Winata, Bryan Wilie, Karissa Vincentio, Xiaohong Li, Adhiguna Kuncoro, Sebastian Ruder, Zhi Yuan Lim, Syafri Bahar, Masayu Khodra, Ayu Purwarianti, and Pascale Fung. 2021. IndoNLG: Benchmark and resources for evaluating Indonesian natural language generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8875–8898, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yong Cao, Min Chen, and Daniel Hershcovich. 2024a. Bridging cultural nuances in dialogue agents through cultural value surveys. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 929–945, St. Julian's, Malta. Association for Computational Linguistics.

Yong Cao, Yova Kementchedjhieva, Ruixiang Cui, Antonia Karamolegkou, Li Zhou, Megan Dare, Lucia Donatelli, and Daniel Hershcovich. 2024b. Cultural Adaptation of Recipes. *Transactions of the Association for Computational Linguistics*, 12:80–99.

Jaemin Cho, Abhay Zala, and Mohit Bansal. 2023. Dall-eval: Probing the reasoning skills and social biases of text-to-image generation models. In *ICCV*.

Colton Clemmer, Junhua Ding, and Yunhe Feng. 2024. Precisedebias: An automatic prompt engineering approach for generative ai to mitigate image demographic biases. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision (WACV)*, pages 8596–8605.

John Corner. 1980. Codes and cultural analysis. *Media, Culture & Society*, 2(1).

Nassim Dehouche. 2021. Implicit stereotypes in pre-trained classifiers. *IEEE Access*, 9:167936–167947.

Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, Dustin Podell, Tim Dockhorn, Zion English, Kyle Lacey, Alex Goodwin, Yannik Marek, and Robin Rombach. 2024. Scaling rectified flow transformers for high-resolution image synthesis. *Preprint*, arXiv:2403.03206.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. 2024. MERA: A comprehensive LLM evaluation in Russian. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9920–9948, Bangkok, Thailand. Association for Computational Linguistics.

Orlando Figes. 2002. *Natasha's dance: A cultural history of Russia*. Macmillan.

Sourojit Ghosh, Pranav Narayanan Venkit, Sanjana Gautam, Shomir Wilson, and Aylin Caliskan. 2024. Do generative ai models output harm while representing non-western cultures: Evidence from a community-centered approach. *arXiv preprint arXiv:2407.14779*.

Mikhail Goloubkov. 2013. Literature and the russian cultural code at the beginning of the 21st century. *Journal of Eurasian Studies*, 4(1):107–113. 20 Years of the Collapse of the Fomer Soviet Union.

Daniel Hershcovich, Stella Frank, Heather Lent, Miryam de Lhoneux, Mostafa Abdou, Stephanie Brandl, Emanuele Bugliarello, Laura Cabello Piqueras, Ilias Chalkidis, Ruixiang Cui, Constanza Fierro, Katerina Margatina, Phillip Rust, and Anders Søgaard. 2022. Challenges and strategies in cross-cultural NLP. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6997–7013, Dublin, Ireland. Association for Computational Linguistics.

Yuichi Inoue, Kento Sasaki, Yuma Ochi, Kazuki Fujii, Kotaro Tanahashi, and Yu Yamaguchi. 2024. Heron-bench: A benchmark for evaluating vision language models in japanese. *Preprint*, arXiv:2404.07824.

Nithish Kannen, Arif Ahmad, Marco Andreetto, Vinodkumar Prabhakaran, Utsav Prabhu, Adji Bousso Dieng, Pushpak Bhattacharyya, and Shachi Dave. 2024. Beyond aesthetics: Cultural competence in text-to-image models. *Preprint*, arXiv:2407.06863.

Sergey Kastryulin, Artem Konev, Alexander Shishenya, Eugene Lyapustin, Artem Khurshudov, Alexander Tselousov, Nikita Vinokurov, Denis Kuznedelev, Alexander Markovich, Grigoriy Livshits, Alexey Kirillov, Anastasiia Tabisheva, Liubov Chubarova, Marina Kaminskaia, Alexander Ustyuzhanin, Artemii Shvetsov, Daniil Shlenskii, Valerii Startsev, Dmitrii Kornilov, Mikhail Romanov, Artem Babenko, Sergei Ovcharenko, and Valentin Khrulkov. 2024. Yaart: Yet another art rendering technology. *Preprint*, arXiv:2404.05666.

Eunsu Kim, Juyoung Suk, Philhoon Oh, Haneul Yoo, James Thorne, and Alice Oh. 2024. CLIcK: A benchmark dataset of cultural and linguistic intelligence in Korean. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3335–3346, Torino, Italia. ELRA and ICCL.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Yulong Liu, Guibo Zhu, Bin Zhu, Qi Song, Guojing Ge, Haoran Chen, GuanHui Qiao, Ru Peng, Lingxiang Wu, and Jinqiao Wang. 2022. Taisu: A 166m large-scale high-quality dataset for chinese vision-language pre-training. In *Advances in Neural Information Processing Systems*, volume 35, pages 16705–16717. Curran Associates, Inc.

Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2024. Stable bias: evaluating societal representations in diffusion models. In *Proceedings of the 37th International Conference on Neural Information Processing Systems*, NIPS '23, Red Hook, NY, USA. Curran Associates Inc.

Midjourney. 2022. Midjourney. `https://www.midjourney.com/`.

Nusrat Jahan Mim, Dipannita Nandi, Sadaf Sumyia Khan, Arundhuti Dey, and Syed Ishtiaque Ahmed. 2024. In-between visuals and visible: The impacts of text-to-image generative ai tools on digital image-making practices in the global south. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, CHI '24, New York, NY, USA. Association for Computing Machinery.

Sourabrata Mukherjee, Atul Kr. Ojha, Akanksha Bansal, Deepak Alok, John P. McCrae, and Ondrej Dusek. 2024. Multilingual text style transfer: Datasets & models for Indian languages. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 494–522, Tokyo, Japan. Association for Computational Linguistics.

Ranjita Naik and Besmira Nushi. 2023. Social biases through the text-to-image generation lens. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 786–808, New York, NY, USA. Association for Computing Machinery.

Shravan Nayak, Kanishk Jain, Rabiul Awal, Siva Reddy, Sjoerd van Steenkiste, Lisa Anne Hendricks, Karolina Stańczak, and Aishwarya Agrawal. 2024. Benchmarking vision language models for cultural understanding. *Preprint*, arXiv:2407.10920.

Denis Peskov, Viktor Hangya, Jordan Boyd-Graber, and Alexander Fraser. 2021. Adapting entities across languages and cultures. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3725–3750, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Rida Qadri, Renee Shelby, Cynthia L. Bennett, and Emily Denton. 2023. Ai's regimes of representation: A community-centered study of text-to-image models in south asia. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 506–517, New York, NY, USA. Association for Computing Machinery.

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. *Preprint*, arXiv:2103.00020.

David Romero, Chenyang Lyu, Haryo Akbarianto Wibowo, Teresa Lynn, Injy Hamed, Aditya Nanda Kishore, Aishik Mandal, Alina Dragonetti, Artem Abzaliev, Atnafu Lambebo Tonja, Bontu Fufa Balcha, Chenxi Whitehouse, Christian Salamea, Dan John Velasco, David Ifeoluwa Adelani, David Le Meur, Emilio Villa-Cueva, Fajri Koto, Fauzan Farooqui, Frederico Belcavello, Ganzorig Batnasan, Gisela Vallejo, Grainne Caulfield, Guido Ivetta, Haiyue Song, Henok Biadglign Ademtew, Hernán Maina, Holy Lovenia, Israel Abebe Azime, Jan Christian Blaise Cruz, Jay Gala, Jiahui Geng, Jesus-German Ortiz-Barajas, Jinheon Baek, Jocelyn Dunstan, Laura Alonso Alemany, Kumaranage Ravindu Yasas Nagasinghe, Luciana Benotti, Luis Fernando D'Haro, Marcelo Viridiano, Marcos Estecha-Garitagoitia, Maria Camila Buitrago Cabrera, Mario Rodríguez-Cantelar, Mélanie Jouitteau, Mihail Mihaylov, Mohamed Fazli Mohamed Imam, Muhammad Farid Adilazuarda, Munkhjargal Gochoo, Munkh-Erdene Otgonbold, Naome Etori, Olivier Niyomugisha, Paula Mónica Silva, Pranjal Chitale, Raj Dabre, Rendi Chevi, Ruochen Zhang, Ryandito Diandaru, Samuel Cahyawijaya, Santiago Góngora, Soyeong Jeong, Sukannya Purkayastha, Tatsuki Kuribayashi, Thanmay Jayakumar, Tiago Timponi Torrent, Toqeer Ehsan, Vladimir Araujo, Yova Kementchedjhieva, Zara Burzo, Zheng Wei Lim, Zheng Xin Yong, Oana Ignat, Joan Nwatu, Rada Mihalcea, Thamar Solorio, and Alham Fikri Aji. 2024. Cvqa: Culturally-diverse multilingual visual question answering benchmark. *Preprint*, arXiv:2406.05967.

Richard Stites. 1992. *Russian popular culture: Entertainment and society since 1900*. Cambridge University Press.

Lukas Struppek, Dom Hintersdorf, Felix Friedrich, Manuel br, Patrick Schramowski, and Kristian Kersting. 2024. Exploiting cultural biases via homoglyphs in text-to-image synthesis. *J. Artif. Int. Res.*, 78.

Ekaterina Taktasheva, Maxim Bazhukov, Kirill Koncha, Alena Fenogenova, Ekaterina Artemova, and Vladislav Mikhailov. 2024. Rublimp: Russian benchmark of linguistic minimal pairs. *Preprint*, arXiv:2406.19232.

Mor Ventura, Eyal Ben-David, Anna Korhonen, and Roi Reichart. 2023. Navigating cultural chasms: Exploring and unlocking the cultural pov of text-to-image models. *arXiv preprint arXiv:2310.01929*.

Arkhipkin Vladimir, Viacheslav Vasilev, Andrei Filatov, Igor Pavlov, Julia Agafonova, Nikolai Gerasimenko, Anna Averchenkova, Evelina Mironova, Bukashkin Anton, Konstantin Kulikov, Andrey Kuznetsov, and Denis Dimitrov. 2024. Kandinsky 3: Text-to-image synthesis for multifunctional generative framework. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 475–485, Miami, Florida, USA. Association for Computational Linguistics.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, Ben Bariach, Iason Gabriel, Verena Rieser, and William Isaac. 2023. Sociotechnical safety evaluation of generative ai systems. *Preprint*, arXiv:2310.11986.

Haryo Akbarianto Wibowo, Erland Hilman Fuadi, Made Nindyatama Nityasya, Radityo Eko Prasojo, and Alham Fikri Aji. 2023. Copal-id: Indonesian language reasoning with local culture and nuances. *arXiv preprint arXiv:2311.01012*.

Anna Wierzbicka. 2002. Russian cultural scripts: The theory of cultural scripts and its applications. *Ethos*, 30(4):401–432.

Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander J Ratner, Ranjay Krishna, Jiaming Shen, and Chao Zhang. 2023. Large language model as attributed training data generator: A tale of diversity and bias. In *Advances in Neural Information Processing Systems*, volume 36, pages 55734–55784. Curran Associates, Inc.

Qingyu Zhang, Xiaoyu Shen, Ernie Chang, Jidong Ge, and Pengke Chen. 2022. Mdia: A benchmark for multilingual dialogue generation in 46 languages. *arXiv preprint*.

Li Zhou, Antonia Karamolegkou, Wenyu Chen, and Daniel Hershcovich. 2023. Cultural compass: Predicting transfer learning success in offensive language detection with cultural features. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 12684–12702, Singapore. Association for Computational Linguistics.

A Side-by-side evaluation for Kandinsky 3.1 and Midjourney v6

According to the main results of the quality assessment, the YandexART 2 and Kandinsky 3.1 models

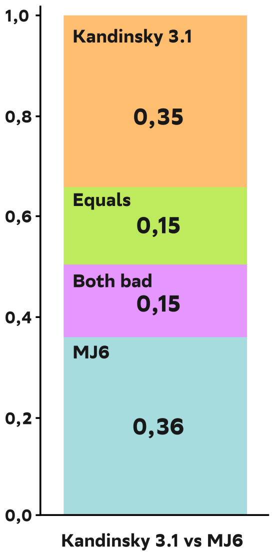
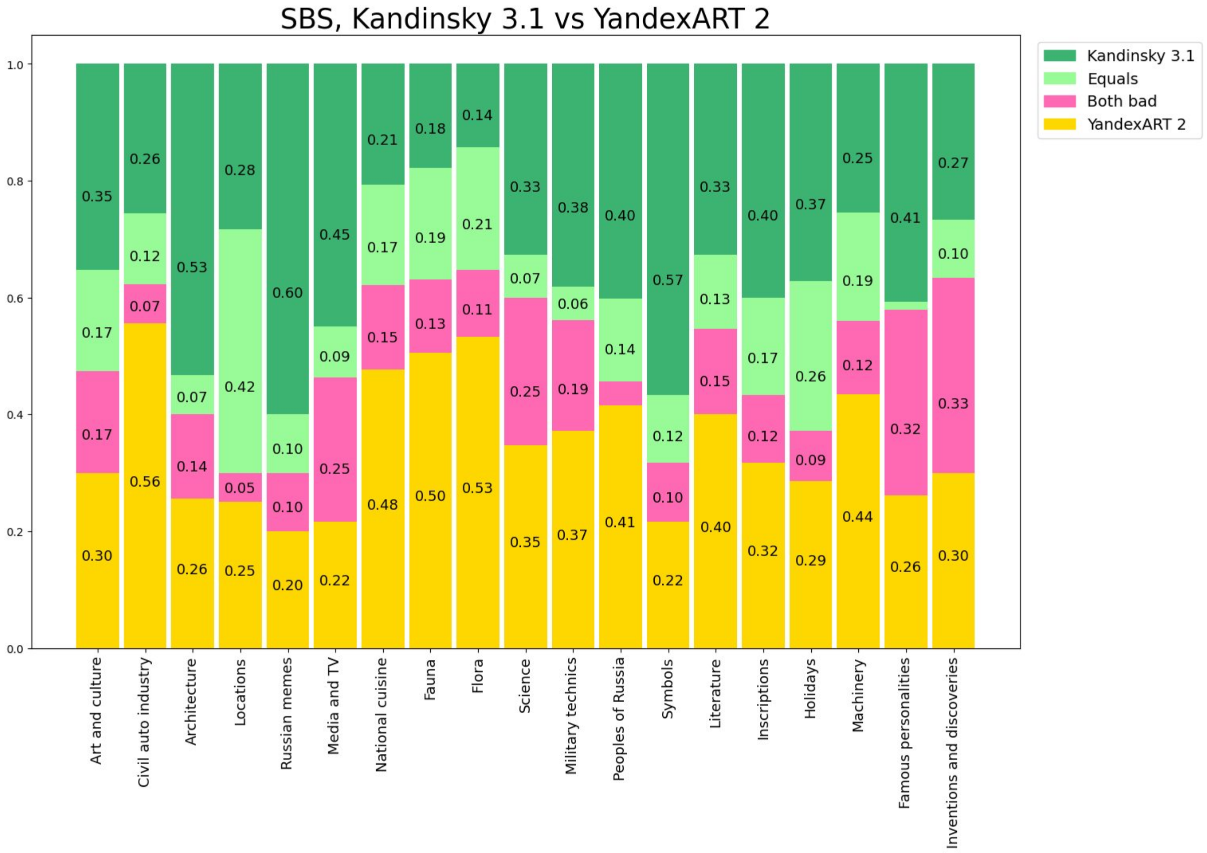
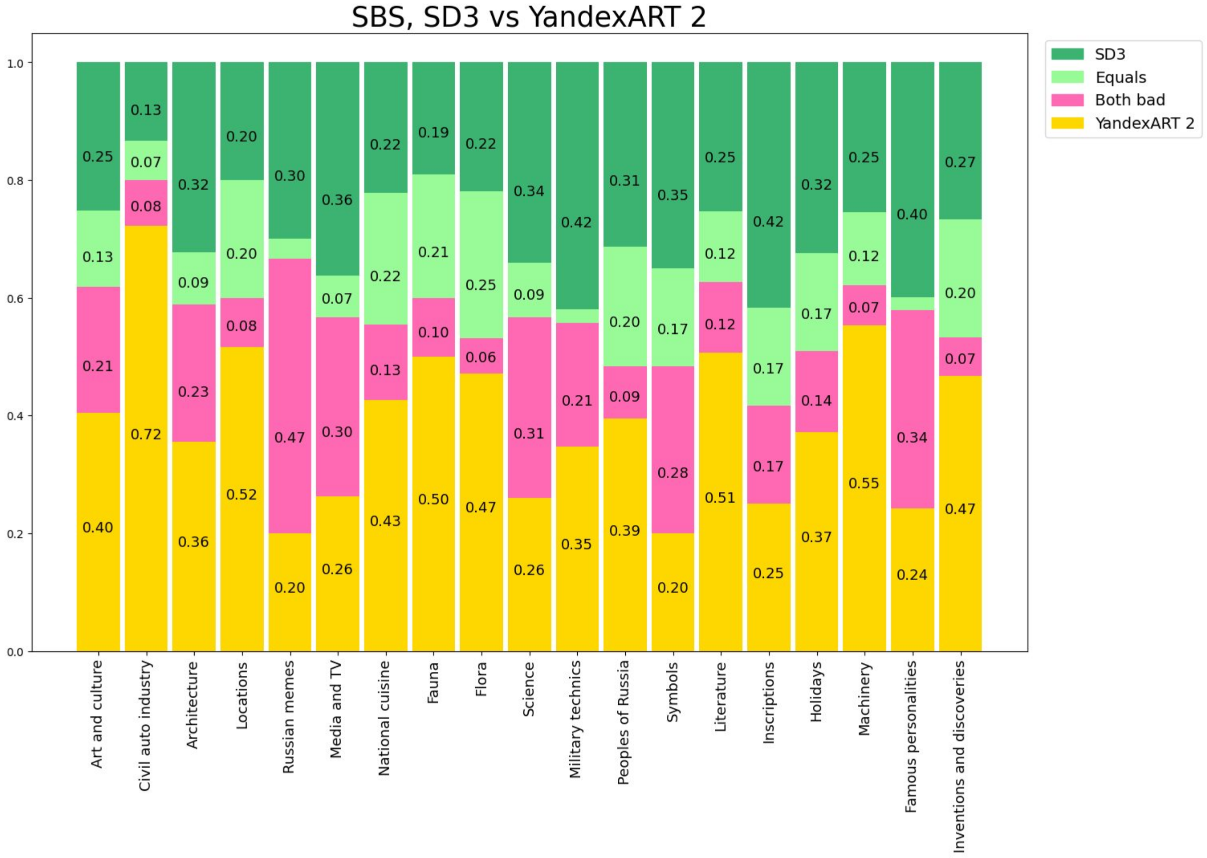


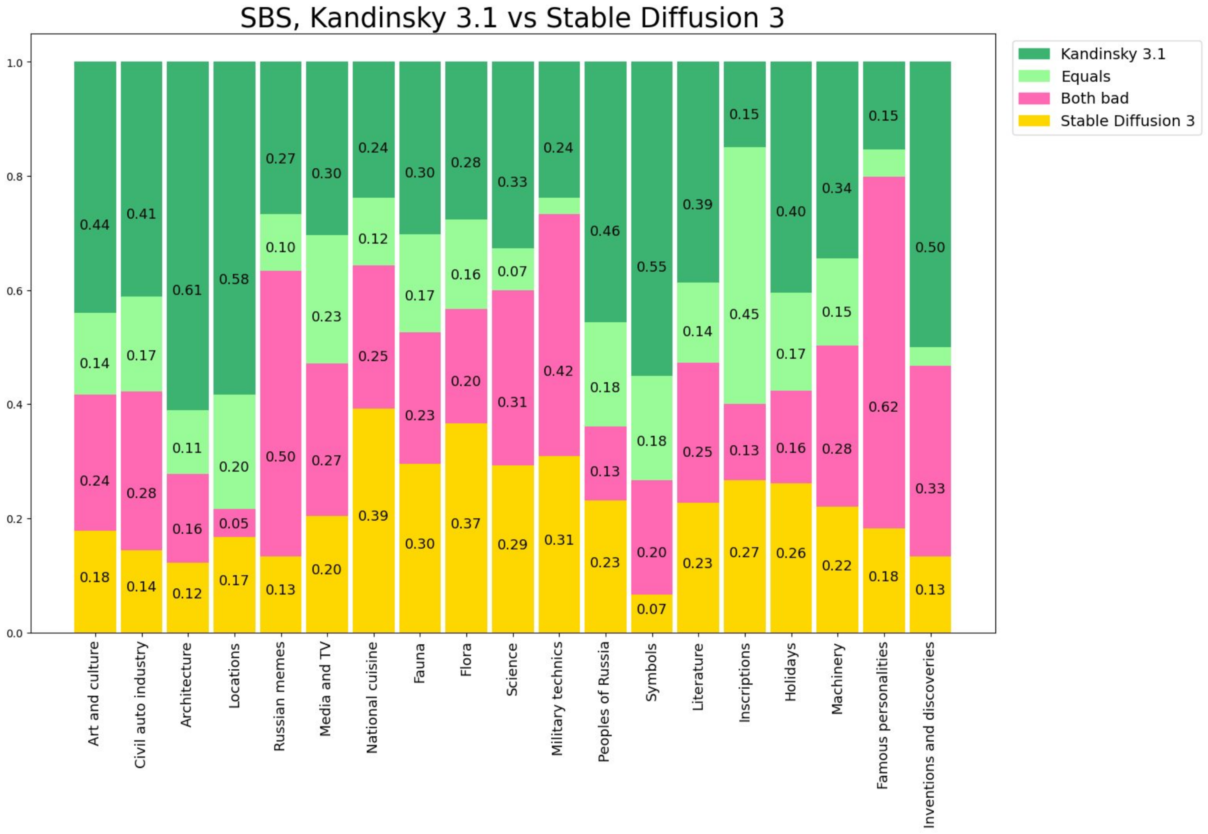
Figure 6: Side-by-side comparison between Kandinsky 3.1 and Midjourney v6.

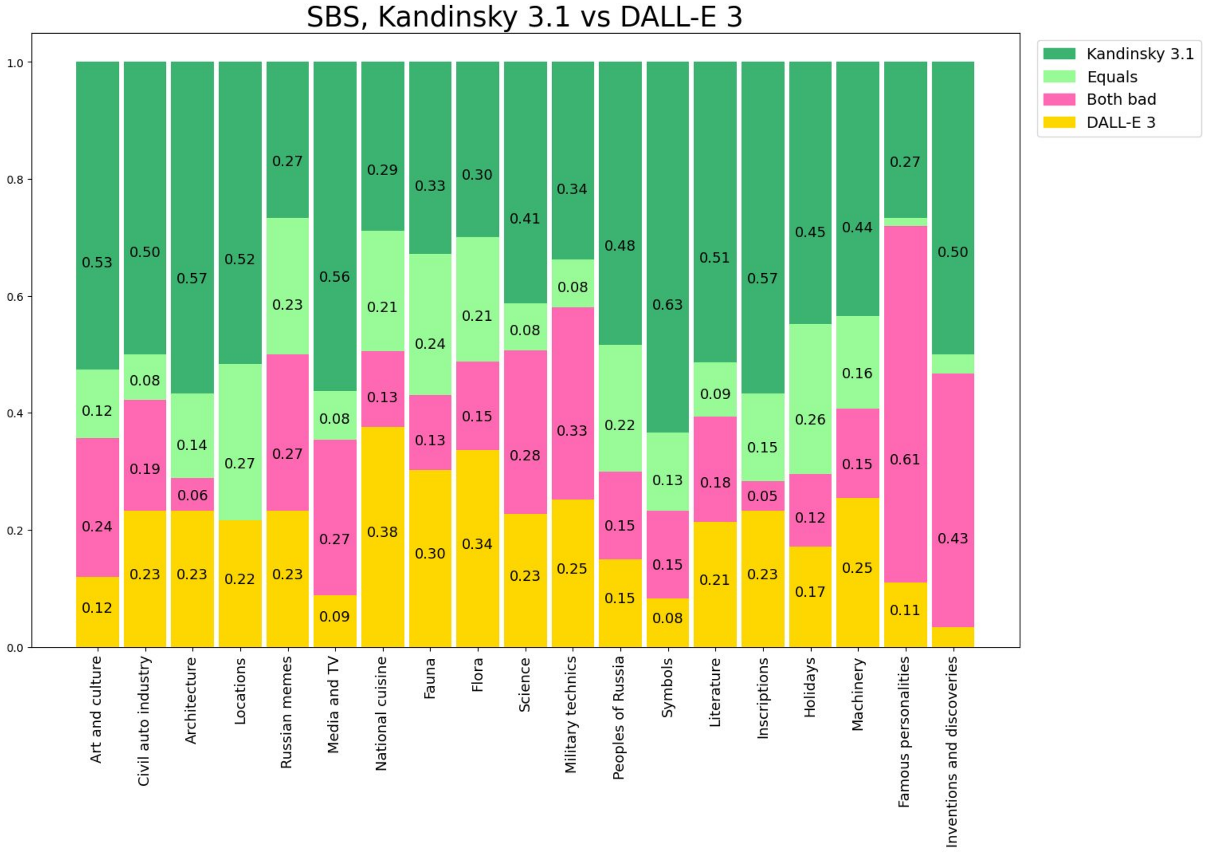
were in the lead, but the YandexART 2 model often censored prompts and did not generate images. For this reason, we chose the Kandinsky 3.1 and additionally compared it with the Midjourney v6 model (Midjourney, 2022). As can be seen from the Figure 6, the models show competitive quality.

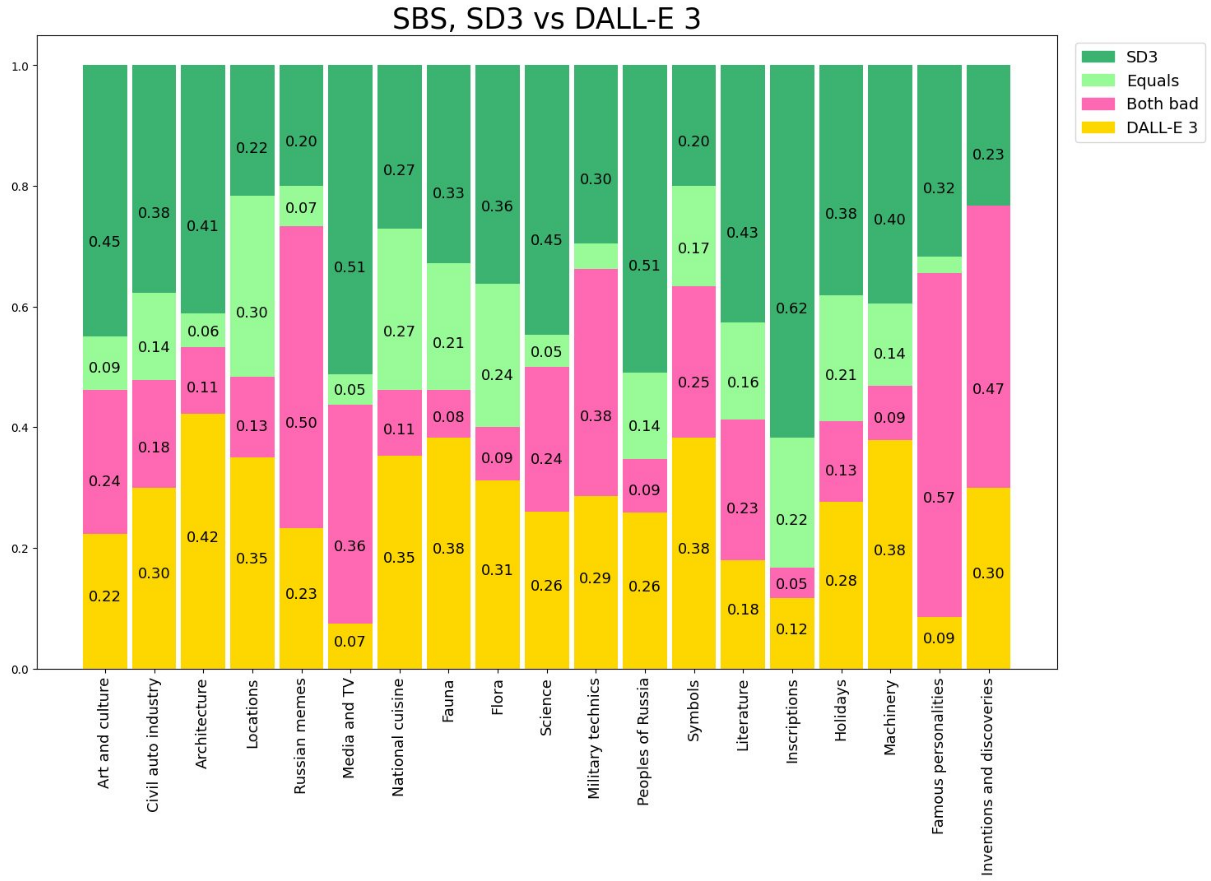
B Side-by-side evaluation by categories

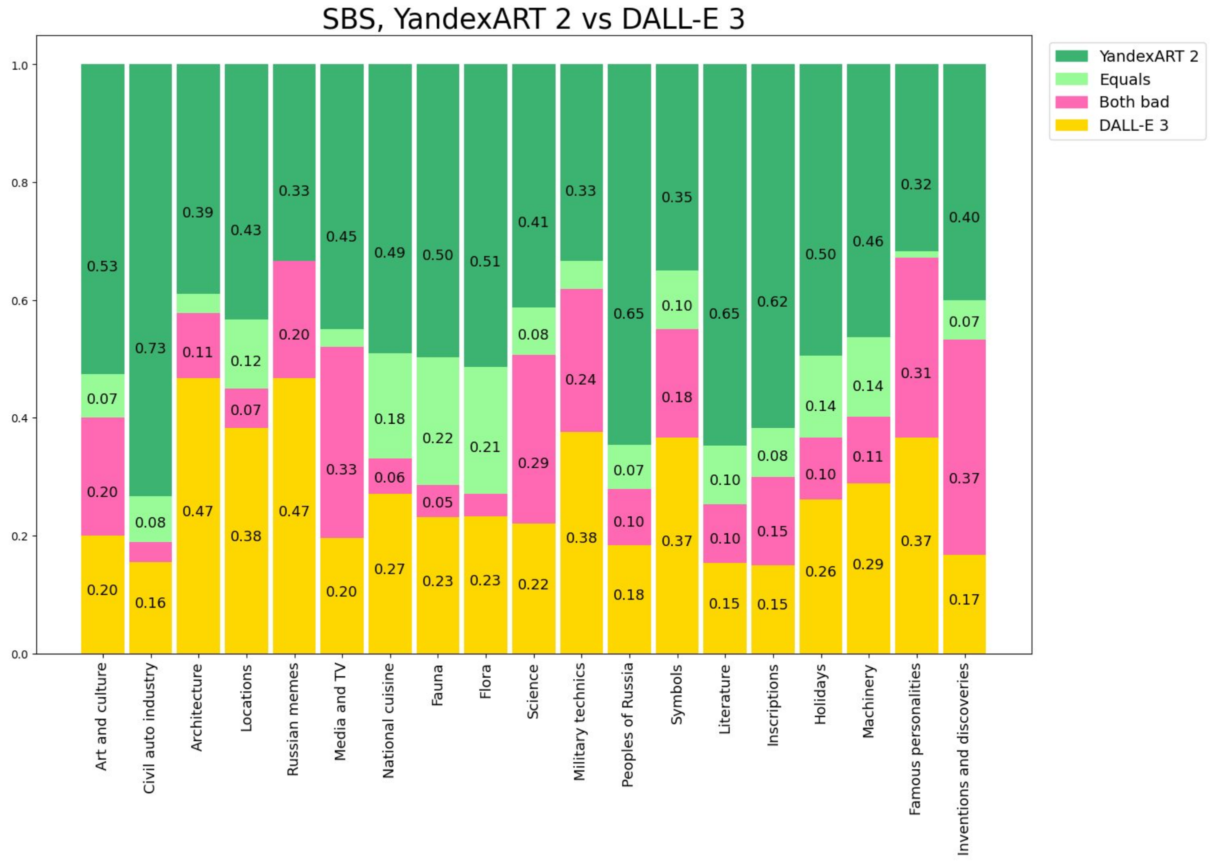

SBS, Kandinsky 3.1 vs YandexART 2
Kandinsky 3.1
Equals
Both bad
YandexART 2
1.0
0.8
0.6
0.4
0.2
0.0
Art and culture: 0.35 0.17 0.17 0.30
Civil auto industry: 0.26 0.12 0.07 0.56
Architecture: 0.53 0.07 0.14 0.26
Locations: 0.28 0.42 0.05 0.25
Russian memes: 0.60 0.10 0.10 0.20
Media and TV: 0.45 0.09 0.25 0.22
National cuisine: 0.21 0.17 0.15 0.48
Fauna: 0.18 0.19 0.13 0.50
Flora: 0.14 0.21 0.11 0.53
Science: 0.33 0.07 0.25 0.35
Military technics: 0.38 0.06 0.19 0.37
Peoples of Russia: 0.40 0.14 0.41
Symbols: 0.57 0.12 0.10 0.22
Literature: 0.33 0.13 0.15 0.40
Inscriptions: 0.40 0.17 0.12 0.32
Holidays: 0.37 0.26 0.09 0.29
Machinery: 0.25 0.19 0.12 0.44
Famous personalities: 0.41 0.32 0.26
Inventions and discoveries: 0.27 0.10 0.33 0.30


SBS, SD3 vs YandexART 2
SD3
Equals
Both bad
YandexART 2
1.0
0.8
0.6
0.4
0.2
0.0
Art and culture: 0.25 0.13 0.21 0.40
Civil auto industry: 0.13 0.07 0.08 0.72
Architecture: 0.32 0.09 0.23 0.36
Locations: 0.20 0.20 0.08 0.52
Russian memes: 0.30 0.47 0.20
Media and TV: 0.36 0.07 0.30 0.26
National cuisine: 0.22 0.22 0.13 0.43
Fauna: 0.19 0.21 0.10 0.50
Flora: 0.22 0.25 0.06 0.47
Science: 0.34 0.09 0.31 0.26
Military technics: 0.42 0.21 0.35
Peoples of Russia: 0.31 0.20 0.09 0.39
Symbols: 0.35 0.17 0.28 0.20
Literature: 0.25 0.12 0.12 0.51
Inscriptions: 0.42 0.17 0.17 0.25
Holidays: 0.32 0.17 0.14 0.37
Machinery: 0.25 0.12 0.07 0.55
Famous personalities: 0.40 0.34 0.24
Inventions and discoveries: 0.27 0.20 0.07 0.47


SBS, Kandinsky 3.1 vs Stable Diffusion 3
Kandinsky 3.1
Equals
Both bad
Stable Diffusion 3
1.0
0.8
0.6
0.4
0.2
0.0
0.44
0.14
0.24
0.18
0.41
0.17
0.28
0.14
0.61
0.11
0.16
0.12
0.58
0.20
0.05
0.17
0.27
0.10
0.50
0.13
0.30
0.23
0.27
0.20
0.24
0.12
0.25
0.39
0.30
0.17
0.23
0.30
0.28
0.16
0.20
0.37
0.33
0.07
0.31
0.29
0.24
0.42
0.31
0.46
0.18
0.13
0.23
0.55
0.18
0.20
0.07
0.39
0.14
0.25
0.23
0.15
0.45
0.13
0.27
0.40
0.17
0.16
0.26
0.34
0.15
0.28
0.22
0.15
0.62
0.18
0.50
0.33
0.13
Art and culture
Civil auto industry
Architecture
Locations
Russian memes
Media and TV
National cuisine
Fauna
Flora
Science
Military technics
Peoples of Russia
Symbols
Literature
Inscriptions
Holidays
Machinery
Famous personalities
Inventions and discoveries


SBS, Kandinsky 3.1 vs DALL-E 3
Kandinsky 3.1
Equals
Both bad
DALL-E 3
1.0
0.8
0.6
0.4
0.2
0.0
0.53
0.12
0.24
0.12
0.50
0.08
0.19
0.23
0.57
0.14
0.06
0.23
0.52
0.27
0.22
0.27
0.23
0.27
0.23
0.56
0.08
0.27
0.09
0.29
0.21
0.13
0.38
0.33
0.24
0.13
0.30
0.30
0.21
0.15
0.34
0.41
0.08
0.28
0.23
0.34
0.08
0.33
0.25
0.48
0.22
0.15
0.15
0.63
0.13
0.15
0.08
0.51
0.09
0.18
0.21
0.57
0.15
0.05
0.23
0.45
0.26
0.12
0.17
0.44
0.16
0.15
0.25
0.27
0.61
0.11
0.50
0.43
Art and culture
Civil auto industry
Architecture
Locations
Russian memes
Media and TV
National cuisine
Fauna
Flora
Science
Military technics
Peoples of Russia
Symbols
Literature
Inscriptions
Holidays
Machinery
Famous personalities
Inventions and discoveries


SBS, SD3 vs DALL-E 3
SD3
Equals
Both bad
DALL-E 3
1.0
0.8
0.6
0.4
0.2
0.0
0.45 0.09 0.24 0.22
0.38 0.14 0.18 0.30
0.41 0.06 0.11 0.42
0.22 0.30 0.13 0.35
0.20 0.07 0.50 0.23
0.51 0.05 0.36 0.07
0.27 0.27 0.11 0.35
0.33 0.21 0.08 0.38
0.36 0.24 0.09 0.31
0.45 0.05 0.24 0.26
0.30 0.38 0.29
0.51 0.14 0.09 0.26
0.20 0.17 0.25 0.38
0.43 0.16 0.23 0.18
0.62 0.22 0.05 0.12
0.38 0.21 0.13 0.28
0.40 0.14 0.09 0.38
0.32 0.57 0.09
0.23 0.47 0.30
Art and culture
Civil auto industry
Architecture
Locations
Russian memes
Media and TV
National cuisine
Fauna
Flora
Science
Military technics
Peoples of Russia
Symbols
Literature
Inscriptions
Holidays
Machinery
Famous personalities
Inventions and discoveries


SBS, YandexART 2 vs DALL-E 3
YandexART 2
Equals
Both bad
DALL-E 3
1.0
0.8
0.6
0.4
0.2
0.0
0.53 0.07 0.20 0.20
0.73 0.08 0.16
0.39 0.11 0.47
0.43 0.12 0.07 0.38
0.33 0.20 0.47
0.45 0.33 0.20
0.49 0.18 0.06 0.27
0.50 0.22 0.05 0.23
0.51 0.21 0.23
0.41 0.08 0.29 0.22
0.33 0.24 0.38
0.65 0.07 0.10 0.18
0.35 0.10 0.18 0.37
0.65 0.10 0.10 0.15
0.62 0.08 0.15 0.15
0.50 0.14 0.10 0.26
0.46 0.14 0.11 0.29
0.32 0.31 0.37
0.40 0.07 0.37 0.17
Art and culture
Civil auto industry
Architecture
Locations
Russian memes
Media and TV
National cuisine
Fauna
Flora
Science
Military technics
Peoples of Russia
Symbols
Literature
Inscriptions
Holidays
Machinery
Famous personalities
Inventions and discoveries