
000 HOW CONFIDENT ARE VIDEO MODELS? 001 002 003 004 005 006 007 008 009 010 011 012 013 014 015 016 017 018 019 020 021 022 023 024 025 026 027 028 029 030 031 032 033 034 035 036 037 038 039 040 041 042 043 044 045 046 047 048 049 050 051 052 053 054 055 056 057 058 059 060 061 062 063 064 065 066 067 068 069 070 071 072 073 074 075 076 077 078 079 080 081 082 083 084 085 086 087 088 089 090 091 092 093 094 095 096 097 098 099 100 101 102 103 104 105 106 107 108 109 110 111 112 113 114 115 116 117 118 119 120 121 122 123 124 125 126 127 128 129 130 131 132 133 134 135 136 137 138 139 140 141 142 143 144 145 146 147 148 149 150 151 152 153 154 155 156 157 158 159 160 161 162 163 164 165 166 167 168 169 170 171 172 173 174 175 176 177 178 179 180 181 182 183 184 185 186 187 188 189 190 191 192 193 194 195 196 197 198 199 200 201 202 203 204 205 206 207 208 209 210 211 212 213 214 215 216 217 218 219 220 221 222 223 224 225 226 227 228 229 230 231 232 233 234 235 236 237 238 239 240 241 242 243 244 245 246 247 248 249 250 251 252 253 254 255 256 257 258 259 260 261 262 263 264 265 266 267 268 269 270 271 272 273 274 275 276 277 278 279 280 281 282 283 284 285 286 287 288 289 290 291 292 293 294 295 296 297 298 299 300 301 302 303 304 305 306 307 308 309 310 311 312 313 314 315 316 317 318 319 320 321 322 323 324 325 326 327 328 329 330 331 332 333 334 335 336 337 338 339 340 341 342 343 344 345 346 347 348 349 350 351 352 353 354 355 356 357 358 359 360 361 362 363 364 365 366 367 368 369 370 371 372 373 374 375 376 377 378 379 380 381 382 383 384 385 386 387 388 389 390 391 392 393 394 395 396 397 398 399 400 401 402 403 404 405 406 407 408 409 410 411 412 413 414 415 416 417 418 419 420 421 422 423 424 425 426 427 428 429 430 431 432 433 434 435 436 437 438 439 440 441 442 443 444 445 446 447 448 449 450 451 452 453 454 455 456 457 458 459 460 461 462 463 464 465 466 467 468 469 470 471 472 473 474 475 476 477 478 479 480 481 482 483 484 485 486 487 488 489 490 491 492 493 494 495 496 497 498 499 500 501 502 503 504 505 506 507 508 509 510 511 512 513 514 515 516 517 518 519 520 521 522 523 524 525 526 527 528 529 530 531 532 533 534 535 536 537 538 539 540 541 542 543 544 545 546 547 548 549 550 551 552 553 554 555 556 557 558 559 559 560 561 562 563 564 565 566 567 568 569 569 570 571 572 573 574 575 576 577 578 579 579 580 581 582 583 584 585 586 587 588 589 589 590 591 592 593 594 595 596 597 598 599 599 600 601 602 603 604 605 606 607 608 609 609 610 611 612 613 614 615 616 617 618 619 619 620 621 622 623 624 625 626 627 628 629 629 630 631 632 633 634 635 636 637 638 639 639 640 641 642 643 644 645 646 647 648 649 649 650 651 652 653 654 655 656 657 658 659 659 660 661 662 663 664 665 666 667 668 669 669 670 671 672 673 674 675 676 677 678 679 679 680 681 682 683 684 685 686 687 688 689 689 690 691 692 693 694 695 696 697 698 699 699 700 701 702 703 704 705 706 707 708 709 709 710 711 712 713 714 715 716 717 718 719 719 720 721 722 723 724 725 726 727 728 729 729 730 731 732 733 734 735 736 737 738 739 739 740 741 742 743 744 745 746 747 748 749 749 750 751 752 753 754 755 756 757 758 759 759 760 761 762 763 764 765 766 767 768 769 769 770 771 772 773 774 775 776 777 778 779 779 780 781 782 783 784 785 786 787 788 789 789 790 791 792 793 794 795 796 797 798 799 799 800 801 802 803 804 805 806 807 808 809 809 810 811 812 813 814 815 816 817 818 819 819 820 821 822 823 824 825 826 827 828 829 829 830 831 832 833 834 835 836 837 838 839 839 840 841 842 843 844 845 846 847 848 849 849 850 851 852 853 854 855 856 857 858 859 859 860 861 862 863 864 865 866 867 868 869 869 870 871 872 873 874 875 876 877 878 879 879 880 881 882 883 884 885 886 887 888 889 889 890 891 892 893 894 895 896 897 898 899 900 901 902 903 904 905 906 907 908 909 909 910 911 912 913 914 915 916 917 918 919 919 920 921 922 923 924 925 926 927 928 929 929 930 931 932 933 934 935 936 937 938 939 939 940 941 942 943 944 945 946 947 948 949 949 950 951 952 953 954 955 956 957 958 959 959 960 961 962 963 964 965 966 967 968 969 969 970 971 972 973 974 975 976 977 978 979 979 980 981 982 983 984 985 986 987 988 989 989 990 991 992 993 994 995 996 997 998 999 1000 1001 1002 1003 1004 1005 1006 1007 1008 1009 1009 1010 1011 1012 1013 1014 1015 1016 1017 1018 1019 1019 1020 1021 1022 1023 1024 1025 1026 1027 1028 1029 1029 1030 1031 1032 1033 1034 1035 1036 1037 1038 1039 1039 1040 1041 1042 1043 1044 1045 1046 1047 1048 1049 1049 1050 1051 1052 1053 1054 1055 1056 1057 1058 1059 1059 1060 1061 1062 1063 1064 1065 1066 1067 1068 1069 1069 1070 1071 1072 1073 1074 1075 1076 1077 1078 1079 1079 1080 1081 1082 1083 1084 1085 1086 1087 1088 1089 1089 1090 1091 1092 1093 1094 1095 1096 1097 1098 1099 1099 1100 1101 1102 1103 1104 1105 1106 1107 1108 1109 1109 1110 1111 1112 1113 1114 1115 1116 1117 1118 1119 1119 1120 1121 1122 1123 1124 1125 1126 1127 1128 1129 1129 1130 1131 1132 1133 1134 1135 1136 1137 1138 1139 1139 1140 1141 1142 1143 1144 1145 1146 1147 1148 1149 1149 1150 1151 1152 1153 1154 1155 1156 1157 1158 1159 1159 1160 1161 1162 1163 1164 1165 1166 1167 1168 1169 1169 1170 1171 1172 1173 1174 1175 1176 1177 1178 1179 1179 1180 1181 1182 1183 1184 1185 1186 1187 1188 1189 1189 1190 1191 1192 1193 1194 1195 1196 1197 1198 1199 1199 1200 1201 1202 1203 1204 1205 1206 1207 1208 1209 1209 1210 1211 1212 1213 1214 1215 1216 1217 1218 1219 1219 1220 1221 1222 1223 1224 1225 1226 1227 1228 1229 1229 1230 1231 1232 1233 1234 1235 1236 1237 1238 1239 1239 1240 1241 1242 1243 1244 1245 1246 1247 1248 1249 1249 1250 1251 1252 1253 1254 1255 1256 1257 1258 1259 1259 1260 1261 1262 1263 1264 1265 1266 1267 1268 1269 1269 1270 1271 1272 1273 1274 1275 1276 1277 1278 1279 1279 1280 1281 1282 1283 1284 1285 1286 1287 1288 1289 1289 1290 1291 1292 1293 1294 1295 1296 1297 1298 1299 1299 1300 1301 1302 1303 1304 1305 1306 1307 1308 1309 1309 1310 1311 1312 1313 1314 1315 1316 1317 1318 1319 1319 1320 1321 1322 1323 1324 1325 1326 1327 1328 1329 1329 1330 1331 1332 1333 1334 1335 1336 1337 1338 1339 1339 1340 1341 1342 1343 1344 1345 1346 1347 1348 1349 1349 1350 1351 1352 1353 1354 1355 1356 1357 1358 1359 1359 1360 1361 1362 1363 1364 1365 1366 1367 1368 1369 1369 1370 1371 1372 1373 1374 1375 1376 1377 1378 1379 1379 1380 1381 1382 1383 1384 1385 1386 1387 1388 1389 1389 1390 1391 1392 1393 1394 1395 1396 1397 1398 1399 1399 1400 1401 1402 1403 1404 1405 1406 1407 1408 1409 1409 1410 1411 1412 1413 1414 1415 1416 1417 1418 1419 1419 1420 1421 1422 1423 1424 1425 1426 1427 1428 1429 1429 1430 1431 1432 1433 1434 1435 1436 1437 1438 1439 1439 1440 1441 1442 1443 1444 1445 1446 1447 1448 1449 1449 1450 1451 1452 1453 1454 1455 1456 1457 1458 1459 1459 1460 1461 1462 1463 1464 1465 1466 1467 1468 1469 1469 1470 1471 1472 1473 1474 1475 1476 1477 1478 1479 1479 1480 1481 1482 1483 1484 1485 1486 1487 1488 1489 1489 1490 1491 1492 1493 1494 1495 1496 1497 1498 1499 1499 1500 1501 1502 1503 1504 1505 1506 1507 1508 1509 1509 1510 1511 1512 1513 1514 1515 1516 1517 1518 1519 1519 1520 1521 1522 1523 1524 1525 1526 1527 1528 1529 1529 1530 1531 1532 1533 1534 1535 1536 1537 1538 1539 1539 1540 1541 1542 1543 1544 1545 1546 1547 1548 1549 1549 1550 1551 1552 1553 1554 1555 1556 1557 1558 1559 1559 1560 1561 1562 1563 1564 1565 1566 1567 1568 1569 1569 1570 1571 1572 1573 1574 1575 1576 1577 1578 1579 1579 1580 1581 1582 1583 1584 1585 1586 1587 1588 1589 1589 1590 1591 1592 1593 1594 1595 1596 1597 1598 1599 1599 1600 1601 1602 1603 1604 1605 1606 1607 1608 1609 1609 1610 1611 1612 1613 1614 1615 1616 1617 1618 1619 1619 1620 1621 1622 1623 1624 1625 1626 1627 1628 1629 1629 1630 1631 1632 1633 1634 1635 1636 1637 1638 1639 1639 1640 1641 1642 1643 1644 1645 1646 1647 1648 1649 1649 1650 1651 1652 1653 1654 1655 1656 1657 1658 1659 1659 1660 1661 1662 1663 1664 1665 1666 1667 1668 1669 1669 1670 1671 1672 1673 1674 1675 1676 1677 1678 1679 1679 1680 1681 1682 1683 1684 1685 1686 1687 1688 1689 1689 1690 1691 1692 1693 1694 1695 1696 1697 1698 1699 1699 1700 1701 1702 1703 1704 1705 1706 1707 1708 1709 1709 1710 1711 1712 1713 1714 1715 1716 1717 1718 1719 1719 1720 1721 1722 1723 1724 1725 1726 1727 1728 1729 1729 1730 1731 1732 1733 1734 1735 1736 1737 1738 1739 1739 1740 1741 1742 1743 1744 1745 1746 1747 1748 1749 1749 1750 1751 1752 1753 1754 1755 1756 1757 1758 1759 1759 1760 1761 1762 1763 1764 1765 1766 1767 1768 1769 1769 1770 1771 1772 1773 1774 1775 1776 1777 1778 1779 1779 1780 1781 1782 1783 1784 1785 1786 1787 1788 1789 1789 1790 1791 1792 1793 1794 1795 1796 1797 1798 1799 1799 1800 1801 1802 1803 1804 1805 1806 1807 1808 1809 1809 1810 1811 1812 1813 1814 1815 1816 1817 1818 1819 1819 1820 1821 1822 1823 1824 1825 1826 1827 1828 1829 1829 1830 1831 1832 1833 1834 1835 1836 1837 1838 1839 1839 1840 1841 1842 1843 1844 1845 1846 1847 1848 1849 1849 1850 1851 1852 1853 1854 1855 1856 1857 1858 1859 1859 1860 1861 1862 1863 1864 1865 1866 1867 1868 1869 1869 1870 1871 1872 1873 1874 1875 1876 1877 1878 1879 1879 1880 1881 1882 1883 1884 1885 1886 1887 1888 1889 1889 1890 1891 1892 1893 1894 1895 1896 1897 1898 1899 1899 1900 1901 1902 1903 1904 1905 1906 1907 1908 1909 1909 1910 1911 1912 1913 1914 1915 1916 1917 1918 1919 1919 1920 1921 1922 1923 1924 1925 1926 1927 1928 1929 1929

054 a crucial safety concern. We illustrate hallucinations in video models in Figure 1. When prompted to
055 generate a video of Jeff Einstein walking on a beach, the video model generates a video of Albert
056 Einstein, an entirely different person, without expressing any doubt in its output. We aim to address
057 this critical challenge by empowering video models to express their uncertainty.

058 Specifically, we propose a framework for uncertainty quantification of video models, consisting of
059 three fundamental components: First, we introduce a *metric* for evaluating the calibration of video
060 models that directly assesses the alignment of the uncertainty estimates with the accuracy of the video
061 generation task. Our metric estimates the rank correlation between uncertainty and task accuracy to
062 measure the calibration error.

063 Second, we derive *S-QUBED* (Semantically-Quantifying Uncertainty with Bayesian Entropy Decom-
064 position), a black-box uncertainty quantification method for video generation models, preserving
065 amenability to the ever-increasing set of closed-source video models. Our key insight is to quantify
066 uncertainty with latent modeling, enabling the rigorous decomposition of predictive uncertainty
067 into its aleatoric and epistemic components. By mapping the input text prompt to a latent space,
068 *S-QUBED* effectively distinguishes between uncertainty arising from ambiguous prompts and uncer-
069 tainty arising from the model’s lack of knowledge. We demonstrate the calibration of *S-QUBED*’s
070 estimates across a broad variety of video generation tasks.

071 Third, we curate a *UQ dataset* of about 40K videos across diverse tasks to facilitate benchmark-
072 ing UQ methods for video models. We generate the data using the open-source model *Cosmos-
073 Predict2* (NVIDIA et al., 2025). We hope that the dataset drives research on uncertainty quantification
074 of video models.

076 2 RELATED WORK

077 **Uncertainty Quantification in Deep Learning.** Deep neural networks (DNNs) are generally
078 difficult to interpret (Li et al., 2022), motivating the development of UQ methods to examine the
079 trustworthiness of their predictions (Abdar et al., 2021). UQ methods in deep learning can be broadly
080 categorized into: *training-free* and *training-based* methods, which constitute a majority of existing
081 work. Training-free methods estimate uncertainty without modifying the model’s architecture,
082 training algorithm, or dataset, e.g., via perturbation techniques (Liu et al., 2024), dropout injection
083 (Loquercio et al., 2020; Ledda et al., 2023), and test-time data augmentation (Ayhan & Berens,
084 2018; Wu & Williamson, 2024). In contrast, training-based methods impose specific architectural
085 design choices to enable uncertainty quantification using Bayesian Neural Networks (BNN) and can
086 be further classified into three categories: (i) variational inference, (ii) Monte-Carlo Dropout, and
087 (iii) Deep Ensemble methods. Assuming that the parameters (weights) of learned models are random
088 variables, BNN methods Kononenko (1989) apply Bayes’ rule to estimate a posterior distribution over
089 these parameters given a prior distribution. However, the exact application of Bayes’ rule is typically
090 intractable, giving rise to approximation techniques, e.g., variational inference (Zhang et al., 2018a),
091 which approximates the posterior distribution using a parametric distribution; Monte-Carlo Dropout
092 Gal & Ghahramani (2016), which samples from the posterior distribution by zeroing-out some
093 weights; and Deep Ensembles (Lakshminarayanan et al., 2017), which train multiple independent
094 models to represent the posterior distribution. Despite their success, traditional UQ methods in deep
095 learning are computationally expensive, limiting their applications in large generative models, e.g.,
096 large language models (LLMs) and vision-language models (VLMs). UQ methods for LLMs/VLMs
097 generally leverage internal activations of these models, or utilize similarity-based metrics or natural-
098 language inference techniques for more efficient UQ (see (Shorinwa et al., 2025) for a detailed
099 discussion).

100 **Uncertainty Quantification in Generative Image/Video Models.** Unlike DNNs and LLMs, UQ
101 of generative image/video models has been relatively underexplored (Franchi et al., 2025). Prior
102 work (Chan et al., 2024) extends Bayesian UQ techniques to denoising diffusion probabilistic models
103 (DDPMs) in generative image modeling by learning a distribution of weights for the diffusion model,
104 enabling the estimation of epistemic uncertainty through the variance across the model’s predictions.
105 Similarly, other approaches (Berry et al., 2024) train latent diffusion models (diffusion ensembles)
106 for UQ by estimating the mutual information over a distribution of the models’ weights, analogous to
107 deep ensembles. However, these training-based UQ methods are challenging to implement, given
108 that diffusion models often have billions of parameters, creating significant computation overhead

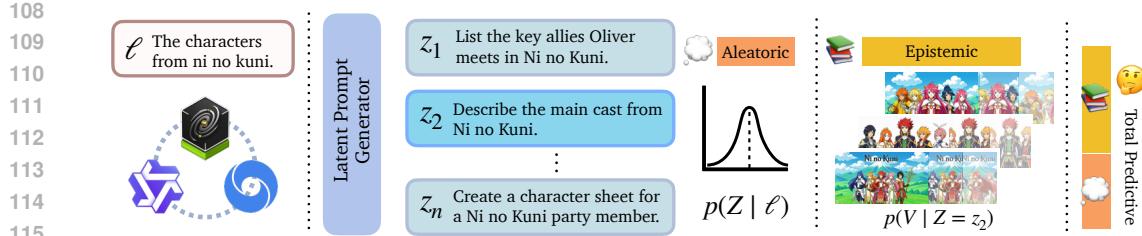


Figure 2: **S-QUBED architecture.** Given a text prompt ℓ , our goal is to quantify the uncertainty of the video generation model. We first generate n latent prompts consistent with ℓ in line with the prompt refinement used by video models, modeling the aleatoric uncertainty as the entropy of the distribution over latent prompts. Then, for each latent prompt, we generate m videos, modeling the epistemic uncertainty as the conditional entropy of the distribution over generated videos. Finally, aggregating the two types of uncertainties yields the total predictive uncertainty.

during training or inference. Drawing insights from black-box UQ methods for LLMs (Manakul et al., 2023; Lin et al., 2023; Becker & Soatto, 2024) which utilize similarity-based techniques for efficient UQ, PUNC (Franchi et al., 2025) explores uncertainty quantification of generative image models in language space. By mapping generated images into language form using a VLM, PUNC leverages widely-used text-based similarity metrics (Zhang et al., 2019; Lin, 2004) to estimate epistemic and aleatoric uncertainty of text-to-image models. Although PUNC addresses the computation limitations of prior UQ methods for generative image models, PUNC is not applicable to [video models](#). To our knowledge, this work is the first exploration of UQ for video world models.

3 PROBLEM FORMULATION

We examine uncertainty quantification of black-box text-conditioned video generation models, which map a text prompt $\ell \in \mathcal{O}$ to a video $v \in \mathcal{V}$ via an unknown stochastic model $f_\theta : \mathcal{T} \mapsto \mathcal{V}$ parametrized by weights θ . Specifically, the video generation process is described by the model:

$$v \sim f_\theta(V \mid \ell), \quad (1)$$

where v is sampled from the conditional distribution f_θ . For an input prompt v , the video generation model has a measure of doubt (uncertainty) associated with the sampled video output v . This uncertainty arises from a variety of sources, e.g., vagueness in the conditioning input ℓ , randomness in the physical evolution of the real-world, limited training data, etc. In this work, we are interested in quantifying the *total* predictive uncertainty associated with v , which can be broadly classified into two categories: *aleatoric* uncertainty and *epistemic* uncertainty.

4 UNCERTAINTY QUANTIFICATION OF GENERATIVE VIDEO MODELS

We present S-QUBED, an efficient method for uncertainty quantification of video generation models, summarized in Figure 2. Without loss of generality, we can decompose the video generation model in Equation (1) using a latent variable $z \in \mathcal{Z}$, modeling the video generation as a two-step process. In the first step, z is sampled from the probability distribution $p(Z \mid \ell)$ conditioned on the input prompt ℓ . In the second step, the video model samples the output video v from the probability distribution $p(V \mid Z = z)$. Note that the application of latent variables is standard in generative modeling, e.g., in variational Bayesian learning (Kingma & Welling, 2013; Sohn et al., 2015; Bhattacharyya et al., 2019), enabling efficient learning and analysis of complex data-generation distributions. Consequently, we can rewrite Equation (1) in the form:

$$f_\theta(V \mid \ell) = \int_{z \in \mathcal{Z}} p(V \mid z, \ell) p(z \mid \ell) dz = \int_{z \in \mathcal{Z}} p(V \mid z) p(z \mid \ell) dz, \quad (2)$$

where we assumed conditional independence of V and ℓ , given z .

Note that the video generation model described by Equation (2) is not limiting. In fact, state-of-the-art text-to-video models refine a user’s prompt using an LLM to generate a much more detailed prompt

162 that is passed into the video generation model. Hence, we can interpret Equation 2 as first sampling
 163 an instance of a fully-specified prompt z from the conditional distribution defined by the input prompt
 164 ℓ , e.g., given the input prompt “a cat doing something,” z may be the more specific prompt “a
 165 cat licking its paws before turning to the camera and meowing...” Subsequently, the video model
 166 generates the output video conditioned on z .

167 **Proposition 1** (Uncertainty Decomposition). *Define the total predictive uncertainty in the output
 168 video as the differential entropy $h(V | \ell)$ of the distribution $f_\theta(V | \ell)$. Then, this quantity can be
 169 decomposed as:*

$$170 \quad h(V | \ell) = h(V | Z) + h(Z | \ell), \quad (3)$$

171 where $h(V | Z)$ represents the epistemic uncertainty in v , and $h(Z | \ell)$ the aleatoric uncertainty.

173 This is a standard decomposition. We provide the proof in Appendix B for completeness. In the rest
 174 of this section, we introduce our approach to estimating these components.

176 4.1 ALEATORIC UNCERTAINTY

178 Aleatoric uncertainty encompasses irreducible randomness from the vagueness (lack of sufficient
 179 specificity) of the conditioning inputs, e.g., “generate a video of a cat doing something.” In video
 180 generation, vagueness in the input prompt increases the randomness of the conditional probability
 181 distribution $p(Z | \ell)$, which is represented by the second term $h(Z | \ell)$ in Equation 3. Note that
 182 $h(Z | \ell)$ is independent of v since the source of uncertainty arises from the input prompt independent
 183 of the second stage of the video generation, e.g., the denoising process in video diffusion models. In
 184 particular, randomness in Z cannot be reduced by training the video model on additional data under
 185 the assumption that we can model $p(Z | \ell)$ almost exactly.

186 As a measure of aleatoric uncertainty, we would expect $h(Z | \ell)$ to be positively correlated with the
 187 vagueness of the input prompt. For example, consider two input prompts: ℓ_1 = “a cat napping” and
 188 ℓ_2 = “a cat doing something”. With ℓ_1 , the pdf of $p(Z | \ell_1)$ will be concentrated on the set:

$$189 \quad \mathcal{A}(\ell_1) = \{“a black cat napping”, “a cat napping on a couch”, “a cat snoring on a couch”, \dots\}. \quad (4)$$

191 However, with ℓ_2 , the pdf of $p(Z | \ell_2)$ will be concentrated on the set:

$$192 \quad \mathcal{A}(\ell_2) = \{“a black cat jumping”, “a cat eating on a couch”, “a cat meowing next to a door”, \dots\}. \quad (5)$$

194 Note that the elements of $\mathcal{A}(\ell_1)$ are more semantically-related (since ℓ_1 is more specific) and are thus
 195 closer in the language (semantic) embedding space compared to elements in $\mathcal{A}(\ell_2)$. Hence, $p(Z | \ell_1)$
 196 will have a lower entropy relative to $p(Z | \ell_2)$.

197 **Modeling the conditional latent distribution.** To compute $h(Z | \ell)$, we need to define a class of
 198 probability distributions that describe the latent-generation process. In this work, we model $p(Z | \ell)$
 199 in a language embedding space using the Von-Mises Fisher (VMF) distribution (Fisher, 1953; Jupp &
 200 Mardia, 1989), drawing insights from prior work (Robertson, 2004; Banerjee et al., 2005; Gopal &
 201 Yang, 2014).

203 The Von-Mises Fisher (VMF) distribution describes a n -dimensional probability distribution on the
 204 $(n - 1)$ -sphere over unit vectors embedded in \mathbb{R}^n , with the probability density function (pdf):

$$205 \quad f_n(x, \mu, \kappa) = C_n(\kappa) \exp(\kappa \mu^\top x), \quad (6)$$

207 with parameters μ and κ denoting the mean direction and concentration parameters, where:

$$208 \quad C_n(\kappa) = \frac{\kappa^{n/2-1}}{(2\pi)^{n/2} I_{n/2-1}(\kappa)}, \quad (7)$$

211 with $I_{n/2-1}$ representing the modified Bessel function of the first kind. The concentration parameter
 212 functions analogously to the inverse variance, providing a measure of the spread of the distribution.

213 We need samples from $p(Z | \ell)$ to fit the VMF distribution. Collecting such data is typically
 214 prohibitively expensive. To overcome this challenge, we leverage LLMs as cost-effective generative
 215 models of $p(Z | \ell)$, noting that video models generally use LLMs to refine prompts prior to generating
 videos.

216 Specifically, given an input prompt ℓ , we generate N *compatible*-but-more-specific prompts from an
 217 LLM. A generated prompt is *compatible* with the input prompt if the generated prompt is consistent
 218 with, i.e., *entails*, the input prompt. However, the converse need not be true: the input prompt might
 219 be underspecified. Subsequently, we compute language embeddings from an embedding model, e.g.,
 220 SentenceFormer (Reimers & Gurevych 2019). Although we could directly fit a VMF to the language
 221 embeddings, we project the language embeddings to a lower-dimensional subspace \mathbb{R}^n using principal
 222 component analysis (PCA) to avoid numerical instability associated with high-dimensional spaces.
 223 We estimate the parameters μ and κ of the VMF distribution in closed-form using approximate
 224 methods (Jupp & Mardia 1989; Sra 2012), circumventing iterative optimization methods.

225 **Estimating Aleatoric Uncertainty.** Given $p(Z | \ell)$, we can compute the aleatoric uncertainty
 226 $h(Z | \ell)$ of v in closed-form via:

$$228 \quad h(Z | \ell) = -\log(C_n(\kappa)) - \frac{\kappa}{\mu_{z|\ell}} \mathbb{E}_Z[Z | \ell](\kappa), \quad (8)$$

230 where $Z \sim \text{VMF}(\mu, \kappa)$ and C_n represents the normalization constant given by Equation (7). The
 231 expected value of the VMF is given by $\mathbb{E}_Z[Z | \ell](\kappa) = W_n(\kappa)\mu_{z|\ell}$, where $W_n = \frac{I_{n/2}(\kappa)}{I_{n/2-1}(\kappa)}$ with the
 232 modified Bessel function of the first kind $I_{n/2}$. We summarize the method for computing aleatoric
 233 uncertainty in Algorithm I.

235 **Algorithm 1:** S-QUBED: Aleatoric Uncertainty Quantification of Generative Video Models

236 **AleatoricUncertainty** (f, ℓ):

$238 \quad \text{Input: Video Model } f, \text{ Input Prompt } \ell;$ $239 \quad \text{Output: Aleatoric Uncertainty } h(Z \ell);$ $240 \quad \mathcal{A}(\ell) \leftarrow \text{Embed(LLM}(\ell)\text{)}; \quad // \text{ Construct } \mathcal{A}(\ell) \text{ from an LLM/VLM}$ $241 \quad \mu_{z \ell}, \kappa_{z \ell} \leftarrow \text{VFit}(\mathcal{A}(\ell)); \quad // \text{ Estimate } p(Z \ell) \text{ with a VMF}$ $242 \quad h(Z \ell) \leftarrow \text{Equation (8)}; \quad // \text{ Compute aleatoric uncertainty } h(Z \ell)$ $243 \quad \text{return } h(Z \ell);$

245 **4.2 EPISTEMIC UNCERTAINTY**

247 Epistemic uncertainty represents the measure of doubt associated with a lack of knowledge, which
 248 generally results from insufficient training data (e.g., Figure I). As a result, epistemic uncertainty is
 249 *reducible* by providing additional training data to the model. In Equation (3), $h(V | Z)$ represents
 250 the epistemic uncertainty of the generated video v , where the uncertainty arises from the limited
 251 knowledge of the video model about concepts expressed by the latent variable $z \in \mathcal{Z}$.

252 For example, consider a video model trained entirely on internet videos of cats and dogs performing
 253 different activities, e.g., running, eating, jumping, meowing/barking. Now, when asked to generate a
 254 video of “a lion roaring in the wild”, the video model might generate different videos across different
 255 runs, with some showing a large cat meowing in a park with significant tree canopy, others showing a
 256 cat making *barking-like* sounds in a forest, etc. Although the generated videos are all conditioned
 257 on semantically-consistent latent variables, the generated videos might be semantically-inconsistent,
 258 since the video model has not been trained on videos of lions. This uncertainty in the generated
 259 videos can be described as *epistemic* and is captured by the entropy term $h(V | Z)$.

260 **Estimating Epistemic Uncertainty.** Note that we can express $h(V | Z)$ in the form:

$$262 \quad h(V | Z) = \mathbb{E}_{z \sim p(z|\ell)}[h(V | Z = z)], \quad (9)$$

263 which can be interpreted as the expected entropy of the distribution of generated videos conditioned on
 264 sampled latent states z from the conditional distribution $p(z | \ell)$. Computing $h(V | Z)$ is challenging
 265 for two reasons: (i) we do not have an explicit model of $p(V | Z = z)$ which is required to compute
 266 $h(V | Z = z)$, and (ii) even with an analytical expression for $p(V | Z = z)$, computing $h(V | Z)$
 267 would require evaluating a double integral, which is intractable in general.

268 To address the first challenge, we approximate the conditional distribution $p(V | Z = z)$ using a
 269 VMF distribution with the parameters μ and κ estimated from samples drawn from the video model.

Likewise, we approximate the expectation in Equation (9) using Monte-Carlo sampling to address the second challenge, which we describe in greater detail.

First, we sample a set of latent variables $\mathcal{E}_{z|\ell}$ conditioned on the input prompt ℓ from the distribution $p(Z | \ell)$, with each $z \in \mathcal{E}_{z|\ell}$ representing specific instances of prompts entailing the input prompt. For each z , we estimate the distribution $p(V | Z = z)$ by generating a set of videos $\mathcal{E}_{v|z}$ from the video model, conditioned on z . Subsequently, we embed these videos with a video embedding model, e.g., S3D (Miech et al., 2020) and fit a VMF to the samples in $\mathcal{E}_{v|z}$. Afterwards, we compute the entropy $h(V | Z = z)$ with:

$$h(V \mid z) = -\log(C_n(\kappa_{v|z})) - \frac{\kappa_{v|z}}{\mu_{v|z}} \mathbb{E}_{v|z}[V \mid Z = z](\kappa_{v|z}), \quad (10)$$

using the estimated VMF parameters $\mu_{v|z}$ and $\kappa_{v|z}$. Finally, we compute an empirical estimate of the expectation of $h(V | Z = z)$ over z sampled from $p(Z | \ell)$. We outline these steps in Algorithm 2.

Algorithm 2: S-QUBED: Epistemic Uncertainty Quantification of Generative Video Models

EpistemicUncertainty (f, ℓ):

```

Input: Video Model  $f$ , Input Prompt  $\ell$ ;  

Output: Epistemic Uncertainty  $h(V \mid z)$ ;  

 $\mathcal{E}_{z|\ell} \leftarrow \text{Embed(LLM}(\ell)\text{)};$  // Construct  $\mathcal{E}_{z|\ell}$  from an LLM/VLM  

foreach  $z \in \mathcal{E}_{z|\ell}$  do  

   $\mathcal{E}_{v|z} \leftarrow \text{Embed}(f(V \mid z))$ ; // Construct  $\mathcal{E}_{v|z}$  from  $f$   

   $\mu_{v|z}, \kappa_{v|z} \leftarrow \text{VFit}(\mathcal{E}_{v|z})$ ; // Estimate  $p(V \mid Z = z)$  from  $f$   

   $h(V \mid Z = z) \leftarrow \text{Equation 10}$ ; // Compute entropy  $h(V \mid Z = z)$   

end  

 $h(V \mid Z) \leftarrow \text{Equation 9}$ ; // Compute epistemic uncertainty  $h(V \mid Z)$   

return  $h(V \mid Z)$ ;

```

5 EXPERIMENTS

We examine the effectiveness of S-QUBED in uncertainty quantification of generative video models, specifically exploring the following questions: (i) *How do we evaluate uncertainty calibration of video models?* (ii) *Are the total predictive uncertainty estimates computed by S-QUBED calibrated?* (iii) *Can S-QUBED effectively estimate both aleatoric and epistemic uncertainty?*

5.1 EVALUATION SETUP

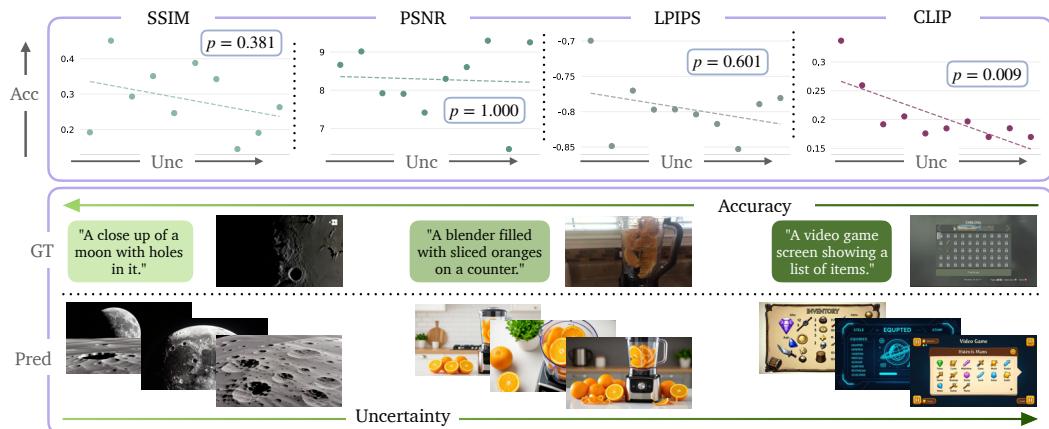
We describe the datasets, models, and metrics used in evaluating our proposed method.

Datasets. We evaluate S-QUBED on two large-scale video generation datasets, VidGen-1M (Tan et al., 2024) and Panda-70M (Chen et al., 2024). Using GPT-5-nano (OpenAI, 2025), we classify the videos in each dataset into five broad categories: animals, food, games, people, and other, a standard approach with video datasets. From each dataset, we subsample about 200 video generation tasks uniformly from each category for evaluation. To address issues with missing video data/metadata in some of the datasets, we sample additional videos from other categories, minimally changing the uniformity of the evaluation dataset.

Implementation. We evaluate S-QUBED on the Cosmos-Predict2 video model (NVIDIA Cosmos, 2025) using the official implementation, which utilizes a text-to-image-to-video pipeline for text conditioning that generates an image from a text prompt, which is used as input to an image-to-video model. We implement our proposed method by sampling 10 latent states, $z_{1:10} \sim p(Z|\ell)$, and subsequently 10 generated videos per latent state, $v_{1:10}^i \sim p(v_i^i|Z = z_i)$. We explore alternative generative video models, including CogVideoX (Yang et al., 2024), Veo 3 (DeepMind, 2025), and OpenSora (Peng et al., 2025) in Appendix C covering a wide range of open-source and closed-source models. However, due to practical limitations on the number of permissible generation requests or prohibitive compute cost, we explore different hyperparameters to enable effective implementations with generation or cost constraints.

324 5.2 HOW DO WE EVALUATE UNCERTAINTY CALIBRATION OF VIDEO MODELS?

326 Uncertainty calibration of video generation models has been underexplored, evidenced by the lack of
 327 purpose-specific calibration metrics. Widely-used calibration metrics, such as the expected calibration
 328 error (ECE) and maximum calibration metrics (MCE) apply only to evaluation settings with discrete
 329 ground-truth answers and errors, e.g., with multiple-choice questions, making them unsuitable in
 330 video generation tasks with real-valued task errors. Consequently, we propose appropriate metrics
 331 for evaluating the calibration of the uncertainty estimates of video models. Specifically, we examine
 332 the Kendall rank correlation (Kendall’s τ) (Kendall, 1938) between the video model’s uncertainty
 333 estimates and an applicable accuracy metric, which captures the degree of monotonicity between
 334 uncertainty and accuracy. We do not utilize Pearson’s rank correlation (Galton, 1895) due to its
 335 assumptions of linearity and normally-distributed data and likewise do not use the Spearman’s rank
 336 correlation coefficient (Spearman, 1987) due to its high sensitivity to outliers.



352 **Figure 3: Calibration Metrics for Video Models.** *Top:* We examine the statistical significance of
 353 the Kendall rank correlation between uncertainty and widely-used perceptual metrics. We find that
 354 the CLIP cosine similarity score provides the most significant correlation. *Bottom:* With the CLIP
 355 accuracy metric, we observe that low human-annotated uncertainty corresponds to smaller variance
 356 in the generated videos and greater accuracy with respect to the ground-truth video. As uncertainty
 357 increases, video prediction accuracy decreases.

358 To compute the rank correlation coefficient, we use the SSIM, PSNR, LPIPS, and CLIP score metrics.
 359 To identify the best metric for assessing calibration, we select 10 generation tasks from the Panda-
 360 70M datasets and rank the tasks in order of increasing uncertainty based on the vagueness of the
 361 text prompt for the task. Note that the vagueness in the prompt directly corresponds to aleatoric
 362 uncertainty, making it an effective proxy measure. Given the human-annotated rankings, we compute
 363 the Kendall rank correlation between uncertainty and each accuracy metric along with a p -value,
 364 which provides a measure of the statistical significance of the correlation. While Panda-70M dataset
 365 consists of tasks with a broad range of descriptive detail from vague to very specific, VidGen-1M
 366 consists of relatively well-detailed tasks. As a result, we do not sample from VidGen-1M, given the
 367 less observable variation in the aleatoric uncertainty. We sample the tasks from Panda-70M dataset to
 368 retain the distribution of instruction detail.

369 We summarize our results in Figure 3. In Panda-70M, the CLIP score metric is strongly negatively
 370 correlated with uncertainty at the 99% significance level. In contrast, the other perceptual metrics
 371 lack a statistically significant correlation with uncertainty. This finding is not entirely surprising,
 372 since CLIP captures semantic information that better reflects the accuracy of the generation task,
 373 unlike the other perceptual metrics which are more susceptible to differences in visual changes.

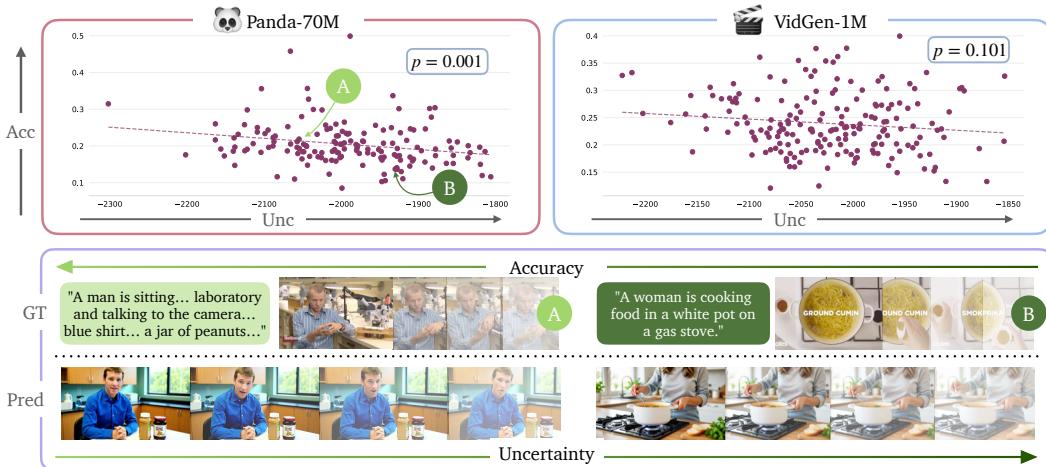
374 Moreover, we visualize the text prompt, ground-truth video, and the first frame of the generated
 375 videos for a few tasks in Figure 3, ranging from low to high uncertainty (rank). We observe that
 376 when uncertainty is low, the model tends to generate very similar videos, which are also close to the
 377 ground-truth, resulting in high accuracy with respect to the CLIP score. As we vary the uncertainty
 378 of the model, we observe greater variance in the generated videos accompanied by notably lower

378 CLIP scores (compared to the other metrics), further demonstrating the utility of the CLIP score as
 379 an accuracy metric.
 380

381 5.3 ARE OUR UNCERTAINTY ESTIMATES CALIBRATED? 382

383 We examine the calibration of our uncertainty estimates in VidGen-1M and Panda-70M, using the
 384 CLIP score accuracy metric given its effectiveness in assessing calibration. We first compute the
 385 total predictive uncertainty associated with each video task using S-QUBED, and then evaluate the
 386 Kendall rank correlation. We define the accuracy of each task as the mean CLIP score across all
 387 generated videos for that task.

388 Figure 4 (left) presents results for Panda-70M. We observe a statistically significant negative correlation
 389 (99% confidence level) between the total uncertainty computed using S-QUBED and the
 390 CLIP score, demonstrating calibration of the uncertainty estimates. The results highlight that as the
 391 uncertainty of the video model decreases, its accuracy increases. Likewise, in VidGen-1M, the total
 392 predictive uncertainty is negatively correlated with the CLIP score at the 89.9% confidence level.
 393 From Figure 4, we see that when the total predictive uncertainty estimates is small (“A”), the video
 394 model generates more accurate videos; in contrast, in tasks with high estimated uncertainty (“B”), the
 395 video model is less accurate.



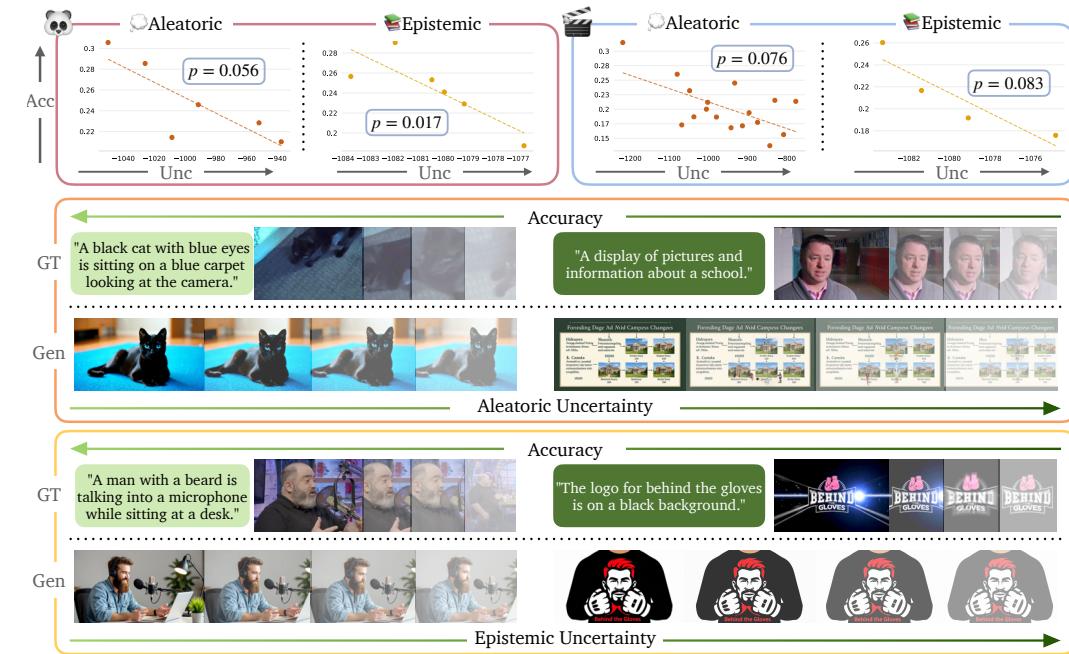
412 Figure 4: **Total Predictive Uncertainty for Video Models.** We assess the calibration of the
 413 total predictive uncertainty computed by S-QUBED. *Top*: correlation between video prediction
 414 accuracy and total uncertainty for Panda-70M and VidGen-1M. We observe a statistically significant
 415 correlation between accuracy and uncertainty for both datasets, signified by the small p -values.
 416 *Bottom*: visualization of two samples from Panda-70M.
 417
 418

419 5.4 CAN S-QUBED EFFECTIVELY ESTIMATE BOTH ALEATORIC AND EPISTEMIC 420 UNCERTAINTY? 421

422 We examine the performance of S-QUBED in decomposing total uncertainty into aleatoric and
 423 epistemic uncertainty. To effectively assess calibration of aleatoric uncertainty, we consider a subset of
 424 each dataset where the epistemic uncertainty is almost zero and compute the rank correlation between
 425 the aleatoric uncertainty of these samples and the CLIP score. Likewise, to evaluate calibration of
 426 epistemic uncertainty, we compute the rank correlation between the epistemic uncertainty and the
 427 CLIP score for samples with relatively zero aleatoric uncertainty. In practice, we select samples with
 428 the lowest aleatoric or epistemic uncertainty, accordingly.

429 In Figure 5, we visualize the Kendall rank correlation between the aleatoric and epistemic uncertainty
 430 and the CLIP score in both datasets. In Panda-70M, we find that aleatoric and epistemic uncertainty are
 431 negatively correlated with accuracy at the 94.5% and 98.3% confidence level. Similarly, in VidGen-1M,
 we observe a statistically significant negative correlation between aleatoric and epistemic

432 uncertainty and the accuracy at the 92.3% and 91.7%, respectively. These results highlight that
 433 S-QUBED can decompose total uncertainty effectively into its aleatoric and epistemic components
 434



456 **Figure 5: Disentangling Aleatoric and Epistemic Uncertainty for Video Models.** We demonstrate
 457 the calibration of the aleatoric uncertainty estimates of S-QUBED in tasks with no epistemic uncer-
 458 tainty, showing statistically significant negative correlation. We do the same for epistemic uncertainty.
 459

460 Further, we visualize text prompts, ground-truth-videos, and generated videos in tasks with low
 461 and high estimated aleatoric uncertainty. We observe that in the low-uncertainty case, the video
 462 model achieves high accuracy, unlike the high-uncertainty case, where the prediction accuracy is
 463 significantly lower. Similarly, we provide some visualizations in the case with low and high estimated
 464 epistemic uncertainty, showing the negative correlation between S-QUBED’s estimated epistemic
 465 uncertainty and video prediction accuracy. Notably, the model does not know the specific “Behind
 466 the Gloves” logo and thus generates a generic logo, unlike predicting the person in the human-centric
 467 videos.

469 6 CONCLUSION

471 We present a framework for empowering video models to express their uncertainty, a critical capability
 472 for safety. Concretely, we introduce a metric for measuring the calibration of UQ methods for video
 473 models and present a calibrated UQ method for video models. Our methods utilize latent modeling to
 474 estimate both aleatoric and epistemic uncertainty, without making any limiting assumptions. Further,
 475 we provide an open-source video dataset for benchmarking UQ methods for video models. Our
 476 experiments demonstrate the calibration of our proposed method and its effectiveness in disentangling
 477 aleatoric and epistemic uncertainty.

479 7 LIMITATIONS AND FUTURE WORK

481 S-QUBED requires generating multiple videos from the video model to estimate epistemic uncertainty,
 482 which poses some computational overhead. Future work will explore more efficient strategies for
 483 sampling videos from the video model, e.g., in the latent space of the video model. Beyond the two
 484 benchmark datasets considered in this work, we will explore extensions to new datasets to augment
 485 the UQ dataset curated for benchmarking calibration. In addition, future work will examine the
 486 application of our method to new open-source models, as they become available.

486 REFERENCES

487

488 Moloud Abdar, Farhad Pourpanah, Sadiq Hussain, Dana Rezazadegan, Li Liu, Mohammad
489 Ghavamzadeh, Paul Fieguth, Xiaochun Cao, Abbas Khosravi, U Rajendra Acharya, et al. A
490 review of uncertainty quantification in deep learning: Techniques, applications and challenges.
491 *Information fusion*, 76:243–297, 2021.

492 Murat Seckin Ayhan and Philipp Berens. Test-time data augmentation for estimation of heteroscedastic
493 aleatoric uncertainty in deep neural networks. In *Medical Imaging with Deep Learning*, 2018.
494

495 Arindam Banerjee, Inderjit S Dhillon, Joydeep Ghosh, Suvrit Sra, and Greg Ridgeway. Clustering on
496 the unit hypersphere using von mises-fisher distributions. *Journal of Machine Learning Research*,
497 6(9), 2005.

498 Evan Becker and Stefano Soatto. Cycles of thought: Measuringilm confidence through stable
499 explanations. *arXiv preprint arXiv:2406.03441*, 2024.
500

501 Lucas Berry, Axel Brando, and David Meger. Shedding light on large generative networks: Estimating
502 epistemic uncertainty in diffusion models. In *The 40th Conference on Uncertainty in Artificial
503 Intelligence*, 2024.

504 Apratim Bhattacharyya, Michael Hanselmann, Mario Fritz, Bernt Schiele, and Christoph-Nikolas
505 Straehle. Conditional flow variational autoencoders for structured sequence prediction. *arXiv
506 preprint arXiv:1908.09008*, 2019.
507

508 Matthew Chan, Maria Molina, and Chris Metzler. Estimating epistemic and aleatoric uncertainty with
509 a single model. *Advances in Neural Information Processing Systems*, 37:109845–109870, 2024.
510

511 Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao,
512 Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, et al. Panda-70m:
513 Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE/CVF
514 Conference on Computer Vision and Pattern Recognition*, pp. 13320–13331, 2024.

515 DeepMind. Veo-3: A text-to-video generation system with audio. Technical Report Tech Report,
516 DeepMind / Google, 2025. Accessed: YYYY-MM-DD.
517

518 Ronald Aylmer Fisher. Dispersion on a sphere. *Proceedings of the Royal Society of London. Series A.
519 Mathematical and Physical Sciences*, 217(1130):295–305, 1953.

520 Gianni Franchi, Nacim Belkhir, Dat Nguyen Trong, Guoxuan Xia, and Andrea Pilzer. Towards
521 understanding and quantifying uncertainty for text-to-image generation. In *Proceedings of the
522 Computer Vision and Pattern Recognition Conference*, pp. 8062–8072, 2025.
523

524 Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model
525 uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059.
526 PMLR, 2016.

527 Francis Galton. Note on regression and correlation. *Proceedings of the Royal Society of London*, 58:
528 240–242, 1895.
529

530 Siddharth Gopal and Yiming Yang. Von mises-fisher clustering models. In *International Conference
531 on Machine Learning*, pp. 154–162. PMLR, 2014.

532 Peter Edmund Jupp and KV Mardia. A unified view of the theory of directional statistics, 1975–1988.
533 *International Statistical Review/Revue Internationale de Statistique*, pp. 261–294, 1989.
534

535 Maurice G Kendall. A new measure of rank correlation. *Biometrika*, 30(1-2):81–93, 1938.
536

537 Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint
538 arXiv:1312.6114*, 2013.

539 Igor Kononenko. Bayesian neural networks. *Biological Cybernetics*, 61(5):361–370, 1989.

540 Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive
541 uncertainty estimation using deep ensembles. *Advances in neural information processing systems*,
542 30, 2017.

543 Emanuele Ledda, Giorgio Fumera, and Fabio Roli. Dropout injection at test time for post hoc
544 uncertainty quantification in neural networks. *Information Sciences*, 645:119356, 2023.

545 Xuhong Li, Haoyi Xiong, Xingjian Li, Xuanyu Wu, Xiao Zhang, Ji Liu, Jiang Bian, and Dejing
546 Dou. Interpretable deep learning: Interpretation, interpretability, trustworthiness, and beyond.
547 *Knowledge and Information Systems*, 64(12):3197–3234, 2022.

548 Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization
549 branches out*, pp. 74–81, 2004.

550 Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantifica-
551 tion for black-box large language models. *arXiv preprint arXiv:2305.19187*, 2023.

552 Yifei Liu, Rex Shen, and Xiaotong Shen. Novel uncertainty quantification through perturbation-
553 assisted sample synthesis. *IEEE transactions on pattern analysis and machine intelligence*, 46(12):
554 7813–7824, 2024.

555 Antonio Loquercio, Mattia Segu, and Davide Scaramuzza. A general framework for uncertainty
556 estimation in deep learning. *IEEE Robotics and Automation Letters*, 5(2):3153–3160, 2020.

557 Potsawee Manakul, Adian Liusie, and Mark JF Gales. Selfcheckgpt: Zero-resource black-box
558 hallucination detection for generative large language models. *arXiv preprint arXiv:2303.08896*,
559 2023.

560 Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zis-
561 serman. End-to-end learning of visual representations from uncurated instructional videos. In
562 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9879–
563 9889, 2020.

564 NVIDIA, :, Niket Agarwal, Arslan Ali, Maciej Bala, Yogesh Balaji, Erik Barker, Tiffany Cai,
565 Prithvijit Chattopadhyay, Yongxin Chen, Yin Cui, Yifan Ding, Daniel Dworakowski, Jiaoqiao Fan,
566 Michele Fenzi, Francesco Ferroni, Sanja Fidler, Dieter Fox, Songwei Ge, Yunhao Ge, Jinwei
567 Gu, Siddharth Gururani, Ethan He, Jiahui Huang, Jacob Huffman, Pooya Jannaty, Jingyi Jin,
568 Seung Wook Kim, Gergely Klár, Grace Lam, Shiyi Lan, Laura Leal-Taixe, Anqi Li, Zhaoshuo
569 Li, Chen-Hsuan Lin, Tsung-Yi Lin, Huan Ling, Ming-Yu Liu, Xian Liu, Alice Luo, Qianli Ma,
570 Hanzi Mao, Kaichun Mo, Arsalan Mousavian, Seungjun Nah, Sriharsha Niverty, David Page,
571 Despoina Paschalidou, Zeeshan Patel, Lindsey Pavao, Morteza Ramezanali, Fitsum Reda, Xiaowei
572 Ren, Vasanth Rao Naik Sabavat, Ed Schmerling, Stella Shi, Bartosz Stefaniak, Shitao Tang, Lyne
573 Tchapmi, Przemek Tredak, Wei-Cheng Tseng, Jibin Varghese, Hao Wang, Haoxiang Wang, Heng
574 Wang, Ting-Chun Wang, Fangyin Wei, Xinyue Wei, Jay Zhangjie Wu, Jiashu Xu, Wei Yang, Lin
575 Yen-Chen, Xiaohui Zeng, Yu Zeng, Jing Zhang, Qinsheng Zhang, Yuxuan Zhang, Qingqing Zhao,
576 and Artur Zolkowski. Cosmos world foundation model platform for physical ai, 2025. URL
577 <https://arxiv.org/abs/2501.03575>.

578 NVIDIA Cosmos. cosmos-predict2: General-purpose world foundation models for physical
579 ai. <https://github.com/nvidia-cosmos/cosmos-predict2>, 2025. Accessed:
580 YYYY-MM-DD.

581 OpenAI. GPT-5 nano, 2025. URL <https://openai.com/gpt-5/>. Large language model.
582 Release date: August 7, 2025.

583 Xiangyu Peng, Zangwei Zheng, Chenhui Shen, Tom Young, Xinying Guo, Binluo Wang, Hang Xu,
584 Hongxin Liu, Mingyan Jiang, Wenjun Li, et al. Open-sora 2.0: Training a commercial-level video
585 generation model in 200k. *arXiv preprint arXiv:2503.09642*, 2025.

586 Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks.
587 In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing
588 and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pp.
589 3982–3992, 2019.

594 Stephen Robertson. Understanding inverse document frequency: on theoretical arguments for idf.
595 *Journal of documentation*, 60(5):503–520, 2004.
596

597 Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z Ren, and Anirudha Majumdar. A survey on
598 uncertainty quantification of large language models: Taxonomy, open research challenges, and future
599 directions. *ACM Computing Surveys*, 2025.

600 Kihyuk Sohn, Honglak Lee, and Xinchen Yan. Learning structured output representation using deep
601 conditional generative models. *Advances in neural information processing systems*, 28, 2015.
602

603 Lin Song, Peter Langfelder, and Steve Horvath. Comparison of co-expression measures: mutual
604 information, correlation, and model based indices. *BMC Bioinformatics*, 13(1):328, 2012. doi:
605 10.1186/1471-2105-13-328.

606 Charles Spearman. The proof and measurement of association between two things. *The American
607 journal of psychology*, 100(3/4):441–471, 1987.
608

609 Suvrit Sra. A short note on parameter approximation for von mises-fisher distributions: and a fast
610 implementation of $\text{is}(\mathbf{x})$. *Computational Statistics*, 27(1):177–190, 2012.

611 Zhiyu Tan, Xiaomeng Yang, Luozheng Qin, and Hao Li. Vidgen-1m: A large-scale dataset for
612 text-to-video generation. *arXiv preprint arXiv:2408.02629*, 2024.
613

614 Team Wan, Ang Wang, Baole Ai, Bin Wen, Chaojie Mao, Chen-Wei Xie, Di Chen, Feiwu Yu, Haiming
615 Zhao, Jianxiao Yang, et al. Wan: Open and advanced large-scale video generative models. *arXiv
616 preprint arXiv:2503.20314*, 2025.

617 Zhou Wang, Alan C Bovik, Hamid R Sheikh, and Eero P Simoncelli. Image quality assessment: from
618 error visibility to structural similarity. *IEEE transactions on image processing*, 13(4):600–612, 2004.
619

620 Luhuan Wu and Sinead A Williamson. Posterior uncertainty quantification in neural networks using data
621 augmentation. In *International Conference on Artificial Intelligence and Statistics*, pp. 3376–3384.
622 PMLR, 2024.

623 Zhuoyi Yang, Jiayan Teng, Wendi Zheng, Ming Ding, Shiyu Huang, Jiazheng Xu, Yuanming Yang,
624 Wenyi Hong, Xiaohan Zhang, Guanyu Feng, et al. Cogvideox: Text-to-video diffusion models with
625 an expert transformer. *arXiv preprint arXiv:2408.06072*, 2024.

626 Cheng Zhang, Judith Bütepape, Hedvig Kjellström, and Stephan Mandt. Advances in variational
627 inference. *IEEE transactions on pattern analysis and machine intelligence*, 41(8):2008–2026, 2018a.
628

629 Richard Zhang, Phillip Isola, Alexei A Efros, Eli Shechtman, and Oliver Wang. The unreasonable
630 effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE conference on
631 computer vision and pattern recognition*, pp. 586–595, 2018b.

632 Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating
633 text generation with bert. *arXiv preprint arXiv:1904.09675*, 2019.
634

635

636

637

638

639

640

641

642

643

644

645

646

647